DIMACS Technical Report 2002-11

Coronary Risk Prediction by Logical Analysis of Data*

by

Sorin Alexe[1]
Eugene Blackstone[2]
Peter L. Hammer[3]
Hemant Ishwaran[2]
Michael S. Lauer[4]
Claire E. Pothier Snader[2]

Acknowledgements

Drs. Lauer, Blackstone, and Ishwaran and Ms. Snader receive support from the National Heart, Lung, and Blood Institute (Grant RO1 HL-66004-1). Dr. Lauer, Dr. Blackstone, and Ms. Snader receive additional support from the American Heart Association (Established Investigator Grant 0040244N). Dr. Hammer and Mr. Alexe receive support from the National Science Foundation (Grant NSF-DMS-9806389) and the Office of Naval Research (Grant N00014-92-J-1375). We gratefully acknowledge all the support which made this study possible.

---

* A preliminary version of this paper was presented at the *Annual Scientific Session of the American College of Cardiology* -- March 18-21, 2001, Orlando, FL and a paper entitled "Use of the "Logical Analysis of Data" Method for Assessing Long-Term Mortality Risk After Exercise Electrocardiography" was submitted for publication to a medical journal.

[1] RUTCOR, Rutgers University, Piscataway, NJ 08854, salexe@rutcor.rutgers.edu
[2] Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195
[3] RUTCOR, Rutgers University, Piscataway, NJ 08854, hammer@rutcor.rutgers.edu
[4] Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195, lauerm@ccf.org

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Research Institute and Telcordia Technologies (formerly Bellcore).

Abstract The objective of this study was to distinguish within a population of patients with known or suspected coronary artery disease groups at high and at low mortality rates. The study was based on Cleveland Clinic Foundation's dataset of 9454 patients, of whom 312 died during an observation period of 9 years. The Logical Analysis of Data method was adapted to handle the disproportioned size of the two groups of patients, and the inseparable character of this dataset -- characteristic to many medical problems. As a result of the study, we have identified a high-risk group of patients representing 1/5 of the population, with a mortality rate 4 times higher than the average, and including 3/4 of the patients who died. The low-risk group identified in the study, representing approximately 4/5 of the population, had a mortality rate 3 times lower than the average.

Keywords: classification, datamining, Logical Analysis of Data, partially defined Boolean functions, risk indices, risk prediction.

AMS classification: 62H30, 06E30, 90C09, 90C10, 90C27, 94C10.

1. Introduction

The dataset of the Cleveland Clinic Foundation (CCF) concerning patients referred for symptoms-exercise electrocardiography between September 1990 and March 1998 was analyzed using a new mathematical method. This method, called Logical Analysis of Data (LAD), is based on combinatorics, optimization and the theory of Boolean functions, and was adapted to handle the type of "inseparable" data which are typical for medical applications. LAD was introduced in 1986 ([9]), the first paper on the topic appeared in 1988 [6], and since then numerous papers (e.g. [1], [2], [7], [8]) have been devoted to various mathematical, computational and applied aspects of this method.

The objective of this study was the construction of a model for distinguishing groups of patients at high and at low mortality risk. Risk stratification is a process common in medical practice by which patients are systematically assessed for the likelihood of developing a poor outcome [3]. For example, in cardiovascular medicine patients with known or suspected coronary artery disease may undergo exercise testing with or without concurrent imaging in order to determine the risk of subsequent death or myocardial infarction [4], [12]-[16], [18]. The purpose of risk stratification is to identify high-risk patients who are most likely to benefit from aggressive therapy and low-risk patients who are best served by conservative care [3], [18].

In the cardiovascular literature, risk stratification schemes are typically based on standard statistical models, such as logistic regression [10] or Cox proportional hazards [5]. A common problem with these approaches is that, although high-risk patients can be easily identified, they usually only account for a minority of subsequent clinical events [17]. Conversely, other risk markers may identify the majority of patients at high risk, but they also include in the same group sizeable numbers of other patients [11]. The ideal risk stratification scheme would identify a small subset of patients who will in fact account for the vast majority of deaths.

As will be seen, an appropriately modified version of LAD produced a model which classifies about 15% of the patients in a category having a mortality rate more than 4 times higher than the average, and about 70% of the patients in a category having a mortality rate of less than 1/5 of the average. Moreover, 2/3 of those patients who died during the observation period appear to be in the high-risk category. Finally, a further refinement of the model increased the number of patients classified into the high-risk or low-risk categories from a total of 85% to more than 97% of the population.

2. The problem and the data

The dataset consisted of observations about 9454 patients, 312 of whom died during the observation period. For each of the patients, 21 variables were recorded, including general data (age, gender), health history (chest pain, hypertension, diabetes, coronary artery disease), medication (beta blockers, verapamil, lipid lowering drugs, aspirin) and specific measurements (resting abnormal ECG, resting heart rate, change in heart rate, chronotropic index, duke treadmill score). Following [4], [12] and [16], Figure 1 presents the complete list of the recorded variables. The analysis took into account all the variables, with the exception of #17 (ttodead), which was only used for cross validation. Variable 12 ("**DEAD"**), indicating whether the patient died during the observation period, was taken as the dependent variable.

	Attribute	Description
1	RestST	resting abnormal ECG (0=no, 1=yes)
2	Pt_age	age in years
3	RBBB	right bundle branch block on ECG (0=no, 1=yes)
4	betablok	use of beta blockers (0=no, 1=yes)
5	dilver	use of diltiazem or verapamil (0=no, 1=yes)
6	lipidrx	lipid lowering drugs (0=no, 1=yes)
7	copd	chronic lung disease (0=no, 1=yes)
8	pvd	peripheral vascular disease (0=no, 1=yes)
9	gender	gender (0=male, 1=female)
10	resthr	resting heart rate in beats per minute
11	aspirin	use of aspirin (0=no, 1=yes)
12	dead	died during followup (0=no, 1=yes)
13	chestp	history of chest pain (0=no, 1=yes)
14	smknow	current smoker (0=no, 1=yes)
15	htn	hypertension (0=no, 1=yes)
16	diabetes	diabetes (0=no, 1=yes)
17	ttodead	length of follow-up in years
18	dhrrec	change in heart rate during the first minute of recovery in beats per minute
19	cri	chronotropic index, a measure of heart rate rise during exercise
20	priorcad	history of known coronary artery disease (0=no, 1=yes)
21	duke	duke treadmill score with typical values between -20 and +15

Figure 1 - Attributes Recorded for Each Patient

For illustration, we present in Figure 2 the data for ten of the surviving patients, and four of the patients who died during the observation period.

Observation	RESTST	PT_AGE	RBBB	BETABLOK	DILVER	LIPIDRX	COPD	PVD	GENDER	RESTHR	ASPIRIN	CHESTP	SMKNOW	HTN	DIABETES	TTODEAD	DHRREC	CRI	PRIORCAD	DUKE	DEAD
1	0	65	0	0	0	0	0	0	0	84	0	0	0	1	0	6.5	27	1.20	0	8.5	**0**
2	0	68	0	0	0	0	0	0	1	77	0	0	0	0	0	8.1	13	1.04	0	9.5	**0**
3	0	63	0	0	0	0	0	0	0	60	0	0	0	0	0	8.4	29	1.01	0	11.5	**0**
4	1	70	0	0	1	0	0	0	0	83	1	0	0	0	1	6.7	8	1.15	1	8.5	**0**
5	0	68	0	0	0	0	0	0	0	71	0	0	0	0	0	8.1	8	0.91	0	6.5	**0**
6	0	60	0	0	1	0	0	0	1	89	1	0	0	0	0	8	24	0.75	1	9.5	**0**
7	0	74	0	0	0	0	0	0	0	85	1	0	0	1	0	7.5	21	1.38	1	9.5	**0**
8	0	79	0	1	1	0	0	0	1	69	0	0	0	1	0	7.5	6	0.47	1	3	**0**
9	1	73	0	0	1	1	0	0	1	89	1	0	0	1	0	8.5	15	0.78	1	-2	**0**
10	0	60	0	0	0	0	0	0	0	61	0	1	0	1	0	2.9	19	0.68	0	8.5	**0**
11	0	66	0	0	0	0	0	0	1	68	0	0	0	1	0	5.3	31	0.86	0	4.5	**1**
12	0	85	0	1	0	0	0	0	1	86	1	0	0	0	0	2.4	13	0.45	1	3	**1**
13	0	75	0	0	1	0	0	0	0	88	0	0	0	0	0	6.2	13	0.96	1	6.5	**1**
14	1	88	0	0	0	0	0	0	0	65	1	0	0	0	1	4.3	7	1.09	1	5.5	**1**

Figure 2 - Sample of Observations

Figure 2 points to a common situation in data analysis. It can be seen from the sample of these 14 observations that none of the variables can by itself make a distinction between those patients who died, and those who did not. Indeed, there is no binary variable taking one of the binary values for all the patients who died, and taking the opposite binary value for all those who survived. Similarly, there is no numerical variable taking only "small" values for one of the two groups of patients, and only "large" values for the other group. On the other hand, it will be seen that there are powerful, robust ways of distinguishing the two groups of data, and that the identification and use of such "classifiers" is the essential feature of LAD.

Let us introduce some terminology. In the discussion below we shall frequently identify a *patient* with an *observation*, and also with the *point in* $\mathbf{R}^{20}$, whose components are the corresponding values of the 20 *variables* (or *attributes*). Such a point will have a positive or a negative label, depending on the value of the *outcome* attribute "**DEAD**"; the observation will be labeled as *positive* if the patient died, and *negative* if he/she survived the observation period.

In previous applications of LAD and other data mining techniques, the $\mathbf{R}^n$ representation of the analyzed datasets generally admitted a more or less "crisp" separation into homogeneous zones, containing only positive or only negative points. An example of such a "separable" dataset is shown in Figure 3, while Figure 4 provides an example for an “inseparable” dataset. An obvious characteristic feature of the inseparable datasets consists in the fact that for many of the given data points are not contained in any “reasonably sized” homogeneous intervals of $\mathbf{R}^n$ (i.e., intervals containing only positive or only negative points); in the example of Figure 2 this property holds for the vast majority of the positive points. The inseparability of the positive and negative data is characteristic to many datasets occurring in medicine, finance and several other important areas of application.

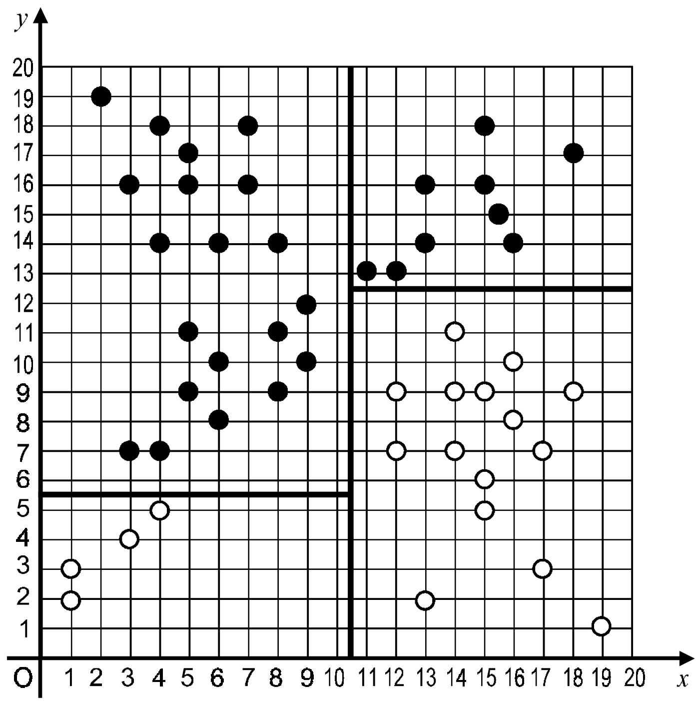

Figure 3. Example of Separable Dataset

● Positive observations ○ Negative observations

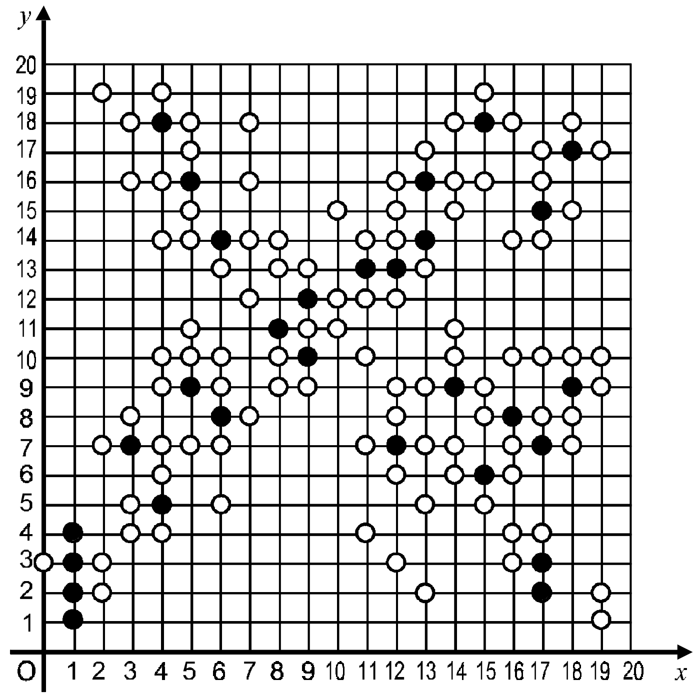

Figure 4. Example of Inseparable Dataset

● Positive observations ○ Negative observations

In order to be able to handle inseparable data we have developed a modified version of the general LAD method, which will be described below and illustrated on the CCF dataset.

3. Patterns

Let us consider a dataset Ω consisting of a finite set of observations, represented as vectors in $\mathbf{R}^n$. We shall denote by Ω^+ and Ω^-, respectively, the subsets of positive and negative observations in Ω.

A subset of points in $\mathbf{R}^n$ identified by upper and/or lower bounds placed on some of their components will be called a *pattern*. More precisely, a *pattern P* is described with the help of two subsets of indices, *I* and *J*, and of a set of real numbers $\alpha_i \left(i \in I\right)$ and $\beta_j \left(j \in J\right)$, called "cutpoints", and is defined as

$$P = \left\{ x \in \mathbf{R}^n \mid x_i \le \alpha_i \left(i \in I\right),\ x_j \ge \beta_j \left(j \in J\right),\ I, J \subseteq \left\{1,2,...,n\right\} \right\}, \qquad (1)$$

A pattern *P* will be called *positive* if $P \cap \Omega^+ \ne \varnothing$ and $P \cap \Omega^- = \varnothing$. Similarly, a pattern *P* will be called *negative* if $P \cap \Omega^- \ne \varnothing$ and $P \cap \Omega^+ = \varnothing$. Clearly, a positive pattern is an interval of $\mathbf{R}^n$ which contains some positive observations and no negative ones. For example, the set of male patients aged 77 or older, having their resting heart rate of at least 97 beats/minute, i.e., those satisfying the conditions

$$\begin{cases} \text{PT\_AGE} \ge 77 \\ \text{GENDER} = 0 \\ \text{RETSTHR} \ge 97, \end{cases} \qquad (2)$$

includes 5 positive observations from the CCF dataset and no negative observations, thus representing a positive pattern.

Positive and negative patterns, introduced in [6], represent in fact one of the fundamental tools used in LAD, and were seen in [2] to give a strong insight into the structure of numerous practical problems of data analysis.

Positive and negative patterns, in their original "pure" form, represent homogeneous intervals in $\mathbf{R}^n$, containing only positive or only negative points. In analyzing the CCF dataset it can be seen that the concepts of positive and negative patterns will have to be extended. One reason for this is the fact that this dataset is quite unbalanced, containing 3.3% positive points (representing patients who died during the observation period), and 96.7% negative points (representing surviving patients). Moreover, the relatively small set of points in $\mathbf{R}^{20}$ representing positive observations is dispersed among the much larger set of points representing negative observations. The presence of *inseparable data* is strongly related to the non-deterministic nature of the relation between the dependent variable and the values of the independent variables. While it is obvious that the fulfillment of various conditions can substantially increase or decrease the *risk of death*, it is equally obvious that the fulfillment of these conditions cannot *guarantee* the occurrence or the absence of this event during the observation period.

As a consequence of the inseparable character of the data, the only positive and negative patterns are of a small size, i.e., contain few data points. For example, the "richest" positive pattern detected in the CCF dataset contains only 11 positive data points, and most of the remaining positive patterns contain not more than 4 or 5 points. Clearly, no conclusions based on such "small" positive patterns can carry much weight. In order to handle this problem, we shall generalize the concept of patterns, allowing beside "pure" positive and negative patterns also those for which the proportion between the positive and negative observations contained in them satisfy some conditions. The requirements to be imposed on these "generalized" patterns (which, for the sake of simplicity, will be simply called "patterns") will be described in the next section.

4. Characteristics of Patterns

The number of patterns which can be detected in a dataset is extremely large. For example, the number of those patterns which can be described using at most three variables in the CCF dataset exceeds 29×10^6. It is therefore important to restrict the attention to reasonably sized subsets of patterns having high information value. In order to extract such subsets, we shall associate to every pattern several parameters, and will limit our search only to those patterns whose associated parameters are below or above some threshold values.

The basic criteria for the experimental selection of parameter-bounds are the following. First, the bounds have to be such that the resulting patterns should be able to provide a powerful distinction between the classes of positive and negative observations. Second, the collection of patterns satisfying the requirements should "cover" the dataset, i.e., each (or at least most of the) observation point(s) should satisfy at least some of the patterns in the collection. Third, the patterns in the collection should be "rich", i.e., the average number of observations satisfying a pattern in the collection should not be too small. Fourth, whenever possible, the resulting patterns should be easily "understandable", i.e., the average number of variables appearing in their expressions ($|I|+|J|$ in the notations of (1)) should be small.

While the parameters can be defined in a rigorous way, and the criteria for imposing limitations on them can also be clearly formulated, the "fine-tuning" of the parameter-bounds is carried out through experimentation.

The three basic parameters to be used in this paper are the concepts of *risk*, *degree* and *prevalence* associated to a somewhat generalized concept of patterns. These concepts along with the associated thresholds used in this analysis will be presented below.

1. *Risk* In order to analyze problems with inseparable data we shall first of all define the concept of *risk* ρ_P of a pattern *P* as the proportion of positive points in the pattern

$$\rho_P = \frac{\left|P \cap \Omega^+\right|}{\left|P \cap \Omega\right|}; \tag{3}$$

here $\left|X\right|$ represents the cardinality of the set *X*. We shall also introduce some real numbers t^+ and t^-, called the *high-risk* and the *low-risk thresholds*, respectively, and shall say that a pattern is of *high risk* (or of *low risk*) if its risk is at least t^+ (respectively, at most t^-).

Consider for example, the patterns P_1 and P_2 defined by

$$P_1 = \begin{cases} RESTST = 0 \\ PT\_AGE \geq 58 \\ RESTHR \geq 94 \end{cases} \tag{4}$$

and

$$P_2 = \begin{cases} RESTST = 1 \\ PT\_AGE \geq 64. \\ GENDER = 0 \end{cases} \tag{5}$$

The defining conditions of pattern P_1 are satisfied by 168 observations, 33 of which are positive. Similarly, the defining conditions of pattern P_2 are satisfied by 347 observations, 54 of which are positive. Clearly, ρ_{P_1} =19.64%, and ρ_{P_2} =15.56%. If, for example, we decide to choose 16.5% as the value of t^+, then P_1 will be viewed as a high-risk pattern, while P_2 will not.

In the case of the CCF dataset, we have identified all patterns for which the mortality rate of the patients satisfying it is at least 5 times larger than the mortality rate of the entire observed population (i.e., it is at least 5x3.3%); therefore we took $t^+ = 16.5\%$. Similarly, we have identified all patterns for which the mortality rate of the patients satisfying it is at most 1/5 of the mortality rate of the entire observed population (i.e., it is at most $\frac{1}{5} \times 3.3\%$); therefore we took $t^- = 0.66\%$.

For further illustration, the pattern

$$P_3 = \begin{cases} \text{PT\_AGE} \geq 68 \\ \text{CHESTP} = 0 \\ \text{DUKE} \leq 5.5 \end{cases} \tag{6}$$

contains 348 patients, of which 59 died during the observation period, i.e., ρ_{P_3} =16.95%, and is therefore a high-risk pattern. Similarly, the pattern

$$P_4 = \begin{cases} 50 \leq \text{PT\_AGE} \leq 52 \\ \text{CRI} \geq 1.05 \end{cases} \tag{7}$$

contains 955 patients, of which 5 died during the observation period, i.e., ρ_{P_4} =0.55%, and is therefore a low-risk pattern.

2. *Degree.* The degree of a pattern is the number of inequalities appearing in its definition (1). For example, both the high-risk pattern P_3 and the low-risk pattern P_4 contain 3 inequalities in their definition. Consequently, each of them is of degree 3.

Usually, we shall concentrate on patterns of low degree, due to two reasons: first their meaning can be easily understood by field experts, and second they can be produced without major computational difficulties.

The analysis of the CCF dataset was based on the use of the high-risk and low-risk patterns of degree at most 3. It should be remarked however, that some high-degree patterns might have a particularly strong information content. For example, the pattern

$$P_5 = \begin{cases} PT\_AGE \leq 59 \\ RESTHR \leq 126 \\ DHRREC \geq 12 \\ DUKE \geq -3.5 \end{cases} \tag{8}$$

of degree 4 contains the remarkable number of 5915 negative points (64.70% of all negative points in the dataset), and contains only 46 positive points, implying that the death rate of patients whose observations satisfy this pattern is of only 0.77%, i.e., less then a quarter of the 3.3% mortality rate of the entire population).

Our experiments with the CCF dataset have shown that in spite of the existence of some higher-degree patterns of remarkably powerful characteristics, confining the analysis to patterns of degree at most 3 leads to results comparable to the analysis of all patterns of degree 4, or even 5. The analysis of all patterns of degree 6 or higher in our dataset is computationally expensive, and not likely to lead to a substantial improvement of the results.

3. *Prevalence*. The number of observations contained in a pattern is called its *absolute prevalence*. The *absolute positive* or *negative prevalence* of a pattern is the number of positive or negative observations contained in it.

In a similar way, the ratio between the absolute prevalence of a pattern and $|\Omega|$ is called the *relative prevalence* of the pattern, while the ratio between the absolute positive (negative) prevalence and $|\Omega^+|$ (respectively $|\Omega^-|$) is called the *relative positive* (respectively *negative*) *prevalence* of the pattern; clearly the absolute negative (positive) prevalence of a positive (respectively negative) pattern is zero.

For illustration, the relative prevalence of the pattern P_4, defined by (7) is $\frac{955}{9454}$, i.e., 10.1%; its relative negative prevalence is $\frac{950}{9142}$, i.e., 10.4%, while its relative positive prevalence is $\frac{5}{312}$, i.e., 1.6%.

The information content of a pattern with a low relative positive and negative prevalence is minimal. Therefore, a significant analysis must be restricted only to patterns having a sufficiently high relative positive or negative prevalence. In this study we shall restrict our attention only to patterns whose relative positive or negative prevalences are of at least 10%.

To summarize, in this study we shall restrict our attention to those high-risk and low-risk patterns, which satisfy the following 3 conditions:

(i) have degree at most 3,

(ii) are defined by high- and low-risk thresholds of 16.5% and 0.66%, respectively,

(iii) have relative positive or negative prevalences of at least 10%.

5. Cutpoints

It has been seen before that patterns are defined by a system of simple inequalities of the form (1). The values of α_i and β_j appearing on the right hand sides of the inequalities (1) are called *cutpoints*.

Originally [6], LAD was developed for the analysis of binary data. When the method was extended to the analysis of numerical data, cutpoints were introduced [1] in order to replace each numerical variable x by several binary variables x_k, which were defined as taking the value 1 if the value of x exceeded some value s_k, and taking the value 0 otherwise. The cutpoints s_k were determined in such a way as to separate as well as possible the positive observations from the negative ones. Although, binary variables do not need cutpoints, for the sake of uniformity we can consider 0.5 to be their only cutpoint. The minimization of the total number of cutpoints to be used was formulated in [1] as a set-covering problem, and it was shown there that most of the natural questions concerning this parameter turn out to be NP-hard, even in the case of only 2 numerical variables.

The determination of cutpoints in the logical analysis of numerical data was based on the assumption that the patterns defined by them will either be positive or negative. In the case of the CCF dataset, as well as in many of the medical, financial, etc datasets, the inseparability of the data renders any attempt to cover the data space with positive and negative patterns of sufficiently large prevalence, futile. Therefore, the logical analysis of this type of datasets has to be based on the use of high-risk and low-risk patterns, rather than positive and negative ones.

A most natural way to identify the cutpoints to be used in the definition of high-risk and low-risk patterns is to use an "equipartitioning" system, which can be built in the following way. Let us assume that for a given dataset the values taken by one of the variables, say x, belong to a set $V = \{v_1, v_2, ..., v_m\}$, where each of the values v_i actually occurs in the dataset, and where $v_1 < v_2 < ... < v_m$. Let further s_i be the midpoint of the interval $[v_i, v_{i+1}]$. The only inequalities $x \leq s$ or $x \geq s$ which will be used in the definition of patterns are those corresponding to the values $s_1, s_2, ..., s_{m-1}$ defined above. The validity of restricting our attention only to these values has been noticed in [1], where additional simplifications were also introduced for the elimination from consideration of some of these values s_i.

In order to restrict further the number of cutpoints in the system $s_1, s_2, ..., s_{m-1}$, we extract from it a subsystem of values $s_{i_1}, s_{i_2}, ..., s_{i_p}$ in such a way that the number of those observations whose component x takes values in anyone of the intervals $[s_q, s_{q+1}]$ should be approximately the same for all $q = 1,2,...,p$.

In the case of the CCF dataset, after comparing the results for p=10, 20, 30, 40, we found the quality of the results to be highest for p=30, and used therefore the corresponding equipartitioning system for the definition of all patterns considered in this study. For illustration we list in Figure 5 the cutpoints for the five numerical variables of the CCF dataset.

	Numerical variables				
Cutpoint	**PT_AGE**	**RESTHR**	**DHRREC**	**CRI**	**DUKE**
1	44.0	58.0	5.0	0.541	-1.5
2	47.0	61.5	8.0	0.614	0.5
3	48.0	63.0	9.0	0.637	1.3
4	49.0	64.5	10.0	0.666	2.3
5	50.0	65.5	10.5	0.704	3.3
6	51.0	67.0	11.5	0.729	3.8
7	52.5	68.5	12.0	0.751	4.0
8	53.5	69.5	13.0	0.771	4.3
9	54.0	70.5	13.5	0.797	4.8
10	55.0	71.0	14.0	0.812	5.0
11	55.5	72.0	14.5	0.826	5.5
12	56.0	73.0	15.0	0.839	5.8
13	57.0	74.0	15.5	0.850	6.0
14	58.0	74.5	16.0	0.863	6.3
15	58.5	75.5	16.5	0.876	6.5
16	59.0	76.4	17.0	0.885	6.5
17	60.0	76.5	17.5	0.898	7.0
18	61.0	78.0	18.0	0.911	7.0
19	61.5	79.0	18.5	0.932	7.5
20	62.0	79.5	19.0	0.940	7.5
21	62.5	80.5	19.5	0.956	7.8
22	63.5	81.5	20.0	0.968	8.0
23	64.5	82.5	21.0	0.984	8.3
24	65.0	84.0	21.5	1.002	8.5
25	66.5	85.5	22.5	1.020	8.8
26	67.5	86.5	23.0	1.039	9.0
27	68.5	88.0	23.5	1.061	9.5
28	70.0	90.5	25.0	1.081	10.0
29	71.5	92.5	26.0	1.104	10.5
30	74.0	98.0	28.0	1.152	11.0

Figure 5. Cutpoints for the 5 Numerical Variables

6. The Pandect

The proposed method requires the selection of a set of cutpoints and the generation of all patterns having prescribed bounds on their degrees, their relative prevalences, and their risk. As mentioned in section 4, in the case of the CCF dataset, these "parameters" were chosen in the following way: the degrees were required not to exceed 3, both the positive and negative relative prevalences were taken as 10%, and the high- and the low-risk thresholds were taken respectively as 16.5% (i.e., five times the mortality rate of the entire population), and 0.66% (i.e., one fifth of the mortality rate of the entire population).

It should be remarked that the collections of intervals defined by other choices of the parameters could be quite different. The actual identification of informative values of the parameters is an experimental process consisting in the creation of various collections of such patterns defined by various parameter values, and the selection of those with the most satisfactory properties.

Taking into account the above definition of the parameters, and using the cutpoint system described in section 5, we have generated all the (approximately 29×10^6) patterns of degree at most 3, and selected from them all those patterns which satisfy our conditions concerning their prevalences and risks. The computer time needed to perform this operation using an Intel III/1GHz processor was of approximately 450 seconds.

Let us denote by Σ the set of all high-risk and low-risk patterns satisfying the prescribed conditions. Σ will be called the *pandect* defined by the chosen system of cutpoints and parameter values. Let further $\Sigma = \Sigma^+ \cup \Sigma^-$, where Σ^+ and Σ^- are the subsets of high-risk and low-risk patterns, respectively.

The pandect Σ constructed for the CCF dataset using the parameter values described before (30 equipartitioning cutpoints/numerical variable, pattern degrees≤ 3, $t^+ = 16.5\%$, $t^- = 0.66\%$, relative positive and negative prevalences of at least 10%) consisted of 783 high-risk and 3940 low-risk patterns.

In order to introduce a quality measure of the pandect, or of any subset *S* of patterns, with $S = S^+ \cup S^-$, where $S^+ \subseteq \Sigma^+$ and $S^- \subseteq \Sigma^-$, let us define the *explained dataset* E_S as being the set consisting of those elements of Ω which satisfy at least one of the patterns in *S*. Similarly, let us define the *unexplained dataset* U_S as being $\Omega - E_S$. One of the quality measures of a system *S* of patterns is the proportion $u_S = \frac{|U_S|}{|\Omega|}$; if the system "explains" the whole set Ω then u_S is zero.

A second important quality measure of *S* is the *overlap* ε_S, defined as

$$\varepsilon_S = \frac{1}{|\Omega|}\left|\left(\bigcup_{\sigma\in S^+}\sigma\right)\cap\left(\bigcup_{\tau\in S^-}\tau\right)\right| \qquad (9).$$

Clearly, the overlap of *S* indicates the proportion of those points of Ω which satisfy some of the high-risk, as well as some of the low-risk patterns; for an ideal system ε_S should be zero.

7. Theories and Models

Clearly, the pandect Σ, containing 4723 high- and low-risk patterns is much too large for practical applications. Moreover, some of the positive and negative observations may be contained in high- or low-risk patterns in Σ, indicating a large redundancy of the pandect. It is natural therefore to investigate nonredundant subsystems of patterns. Obviously, the explained dataset $E_{\Sigma'}$ of such a subsystem cannot be larger than the explained dataset E_Σ of the complete set Σ.

We shall call a subset $\mathbf{T}^+$ of high-risk patterns of Σ^+ a *positive theory* if $E_{T^+} = E_{\Sigma^+}$, i.e.,

$$x \in \Omega^+ \cap \left(\bigcup_{\sigma\in\Sigma^+}\sigma\right) \Rightarrow x \in \bigcup_{\sigma\in T^+}\sigma; \qquad (10)$$

clearly, every observation contained in some high-risk pattern of the pandect is also contained in a high-risk pattern of any positive theory.

Similarly, we shall call a subset $\mathbf{T}^-$ of low-risk patterns of Σ^- a *negative theory* if $E_{T^-} = E_{\Sigma^-}$, i.e.,

$$x \in \Omega^- \cap \left(\bigcup_{\sigma\in\Sigma^-}\sigma\right) \Rightarrow x \in \bigcup_{\sigma\in T^-}\sigma; \qquad (11)$$

clearly, every observation contained in some low-risk pattern of the pandect is also contained in a low-risk pattern of any negative theory.

A pair consisting of a positive and a negative theory will be called a *model*.

It is clear that the minimum sizes of the positive and of the negative theories can be determined by solving separately two naturally associated set-covering problems.

In the case of the CCF dataset we have determined a model **T** consisting of 42 high-risk and 77 low-risk patterns. The model is presented in Figures 7 and 8, where the prevalences and the risks of each pattern are specified, along with the inequalities defining the pattern in terms of the original variables.

The quality of the model can be measured by the size of its unexplained dataset U_T, which consists of 247 points (i.e., u_T =2.61%), and of its overlap ε_T =12.26%.

In order to get a more detailed understanding of the classification power of the model, we shall introduce now the concept of a classification matrix. The *classification matrix* of a model **T** consists of 4 rows (corresponding respectively to positive data **(P)**, negative data **(N), Risk,** and prevalence **(Prev)**), and 4 columns:

1. **HRT** - percentage of observations satisfying high-risk patterns of **T** only,
2. **LRT** - percentage of observations satisfying low-risk patterns of **T** only,
3. **HLT** - percentage of observations satisfying both high and low-risk patterns of **T**,
4. **UT** - percentage of observations not satisfying any of the high or low-risk patterns of **T**.

In an ideal case, the only nonzero element in the row **P** would be the entry 100% appearing in column **HRT**, and similarly the only nonzero entry in the row **N** would be 100% appearing in the column **LRT.** In the same case the first two entries in the row **Risk** would be 100% and 0%, respectively, and the only two nonzero entries in the row **Prev** would appear in the first two entries and would add up to 100%.

For the CCF dataset the model **T** produces the following classification matrix:

	HRT	**LRT**	**HLT**	**UT**
P	63.78%	13.78%	20.83%	1.60%
N	13.17%	72.22%	11.97%	2.65%
Risk	**14.18%**	**0.65%**		
Prev	**14.84%**	**70.28%**	**12.26%**	**2.61%**

Figure 6. Classification Matrix for Model

As an example we show how the risk for the column **HRT** in Figure 6 was calculated. This column includes 63.78% of the 312 positive observations representing patients who died during the observation period (i.e., 199 patients), and 13.17% of the 9142 negative observations (i.e., 1204 patients). Therefore, the risk in the high-risk subset of observations reported in this column is $\frac{199}{199+1204}$, i.e., 14.18%.

In conclusion, the model:

- identifies a set of high-risk observations with a mortality rate of 14.18%, and a set of low-risk observations with a mortality rate of 0.65%,
- classifies correctly and unambiguously 63.78% of the positive observations as being at high-risk and 72.22% of the negative ones as being at low-risk,
- classifies erroneously 13.78% of the positive observations and 13.17% of the negative ones,
- does not provide any classification for 2.61% of the observations.

	Prevalence		Pattern parameters			Pattern description												
Pattern	**Positive**	**Negative**	**Degree**	**Relative positive prevalence**	**Risk**	**RESTST**	**PT_AGE**	**PVD**	**GENDER**	**RESTHR**	**ASPIRIN**	**CHESTP**	**HTN**	**DIABETES**	**DHRREC**	**CRI**	**PRIORCAD**	**DUKE**
P1	38	178	3	12.18%	17.59%	1	>65						1					
P2	50	242	3	16.03%	17.12%	1	>60									<0.823		
P3	39	191	3	12.50%	16.96%	1	>61											<4
P4	35	171	3	11.22%	16.99%	0										<0.719	1	
P5	40	198	3	12.82%	16.81%		>64	0										<1.5
P6	49	241	3	15.71%	16.90%		>57		0	>87								
P7	85	422	3	27.24%	16.77%		>56		0							<0.762		
P8	46	224	2	14.74%	17.04%		>64			>85								
P9	39	181	3	12.50%	17.73%		>57			>87			1					
P10	45	227	3	14.42%	16.54%		>42			>93					<11			
P11	72	361	3	23.08%	16.63%		>55			>71					<9			
P12	61	301	3	19.55%	16.85%		>55			>82					<11			
P13	79	394	3	25.32%	16.70%		>57			>61					<9			
P14	58	280	3	18.59%	17.16%		>60			>82					<13.5			
P15	58	287	3	18.59%	16.81%		>64			>76					<14.5			
P16	45	222	3	14.42%	16.85%		>60			>87						<1.119		
P17	50	250	3	16.03%	16.67%		>63			>76						<0.924		
P18	56	283	3	17.95%	16.52%		>64			>78						<1.097		
P19	46	232	3	14.74%	16.55%		>63			>74								<4
P20	58	289	3	18.59%	16.71%		>64			>76								<6.25
P21	68	341	3	21.79%	16.63%		>52				0					<0.762		
P22	55	272	3	17.63%	16.82%		>64				0					<0.877		
P23	72	359	3	23.08%	16.71%		>64						1			<1.008		
P24	69	349	3	22.12%	16.51%		>52								<9	<0.891		
P25	33	167	3	10.58%	16.50%		>57								<14.5			<1.5
P26	94	452	3	30.13%	17.22%		>61								<11			<9
P27	71	358	3	22.76%	16.55%		>64								<22			<4
P28	87	438	3	27.88%	16.57%		>67								<14.5			<11
P29	82	412	3	26.28%	16.60%		>67								<17.5			<5.5
P30	59	296	3	18.91%	16.62%		>67								<26.5			<3
P31	63	317	3	20.19%	16.58%		>55									<0.762	1	
P32	39	196	3	12.50%	16.60%		>58									<0.924		<1.5
P33	74	373	3	23.72%	16.55%		>61									<0.97		<3
P34	65	323	3	20.83%	16.75%				0						<12			<4.5
P35	45	225	3	14.42%	16.67%					>87					<11	<0.953		
P36	43	208	3	13.78%	17.13%						0				<16.5		1	
P37	48	241	3	15.38%	16.61%						0					<0.659		<11
P38	59	296	3	18.91%	16.62%						0					<0.796		<4.5
P39	32	161	3	10.26%	16.58%						0						1	<4.5
P40	87	416	3	27.88%	17.30%							0			<9			<7.5
P41	37	180	3	11.86%	17.05%										<16.5	<0.944		<1.5
P42	35	174	2	11.22%	16.75%											<0.762		<1.5

Figure 7. Positive Theory Consisting of 42 High-Risk Patterns

	Prevalence		Pattern parameters			Pattern description												
Pattern	**Positive**	**Negative**	**Degree**	**Relative negative prevalenc**	**Risk**	**RESTST**	**PT_AGE**	**PVD**	**GENDER**	**RESTHR**	**ASPIRIN**	**CHESTP**	**HTN**	**DIABETES**	**DHRREC**	**CRI**	**PRIORCAD**	**DUKE**
N1	31	4709	2	51.51%	0.65%	0	<54											
N2	20	3132	1	34.26%	0.63%		<48											
N3	7	1140	2	12.47%	0.61%		<58			<63								
N4	5	935	2	10.23%	0.53%		<58			<62								
N5	6	948	3	10.37%	0.63%		>45, <52			<73								
N6	6	915	3	10.01%	0.65%		>47, <55			<72								
N7	10	1561	2	17.08%	0.64%		<60								>26.5			
N8	20	3062	2	33.49%	0.65%		<64									>1.008		
N9	5	921	3	10.07%	0.54%		>51, <57									>0.924		
N10	30	4516	2	49.40%	0.66%		<52										0	
N11	6	975	3	10.67%	0.61%		>49, <53											>7
N12	6	944	3	10.33%	0.63%		<57		1						>14.5			
N13	4	942	3	10.30%	0.42%		<53		1							>0.762		
N14	3	928	3	10.15%	0.32%		<58		1							>0.862		
N15	4	919	3	10.05%	0.43%		<60		1							>0.902		
N16	5	940	3	10.28%	0.53%		<64		1							>0.913		
N17	5	944	3	10.33%	0.53%		<55		1									>5
N18	6	920	3	10.06%	0.65%		<54			>74, <80								
N19	6	941	3	10.29%	0.63%		<55			>79, <88								
N20	6	926	3	10.13%	0.64%		<58			>64, <69								
N21	6	944	3	10.33%	0.63%		<59			<62				0				
N22	4	917	3	10.03%	0.43%		<60			<67					>24			
N23	6	921	3	10.07%	0.65%		<62			<72					>26.5			
N24	5	922	3	10.09%	0.54%		<58			>80					>17			
N25	4	931	3	10.18%	0.43%		<58			>81					>16			
N26	4	935	3	10.23%	0.43%		<58			>85					>12			
N27	5	929	3	10.16%	0.54%		<60			>79					>19.5			
N28	5	938	3	10.26%	0.53%		<60			>84					>15			
N29	6	923	3	10.10%	0.65%		<62			>81					>17			
N30	6	929	3	10.16%	0.64%		<64			>80					>18			
N31	5	953	3	10.42%	0.52%		<58			<63						>0.844		
N32	5	930	3	10.17%	0.53%		<60			<65						>0.902		
N33	6	935	3	10.23%	0.64%		<68			<67						>0.989		
N34	6	930	3	10.17%	0.64%		<59			>85						>0.924		
N35	6	938	3	10.26%	0.64%		<62			>85						>0.944		
N36	6	916	3	10.02%	0.65%		<65			>81						>1.038		
N37	5	933	3	10.21%	0.53%		<58			>85								>7
N38	5	915	3	10.01%	0.54%		<59								>22.5, <26.5			
N39	6	965	3	10.56%	0.62%		<54								>15			<7.65
N40	5	956	3	10.46%	0.52%		<57								>19.5			<9.5
N41	6	924	3	10.11%	0.65%		<59									>0.934, <0.98		
N42	6	928	3	10.15%	0.64%		<68									>1.05, <1.097		
N43	6	925	3	10.12%	0.64%		<55									>0.902		<8.5
N44	6	934	3	10.22%	0.64%		<60									>0.953		<7.65
N45	11	1837	2	20.09%	0.60%			0										>0
N46	5	963	3	10.53%	0.52%					>81	0							>10.9
N47	7	1116	3	12.21%	0.62%					<68			0					>0
N48	6	923	3	10.10%	0.65%					<67					>9	>0.998		
N49	6	958	3	10.48%	0.62%					<68					>15	>0.989		
N50	6	927	3	10.14%	0.64%					>74					>19.5	>0.989		
N51	6	973	3	10.64%	0.61%					>74					>18	>0.998		
N52	6	927	3	10.14%	0.64%					>74					>20.5	>0.934		
N53	6	925	3	10.12%	0.64%					>75					>19.5	>0.961		
N54	6	937	3	10.25%	0.64%					<69					>22.5			>10.9
N55	6	919	3	10.05%	0.65%					<69					>24.5			>5
N56	6	936	3	10.24%	0.64%					<73					>26.5			>6
N57	9	1417	3	15.50%	0.63%					>65					>21.5			>8.5
N58	9	1360	3	14.88%	0.66%					>71					>19.5			>7.65
N59	6	964	3	10.54%	0.62%					<71						>0.961, <1.05		
N60	6	932	3	10.19%	0.64%					<72						>0.97, <1.05		
N61	6	960	3	10.50%	0.62%					<72						>1.008, <1.097		
N62	6	927	3	10.14%	0.64%					<85						>0.998, <1.038		
N63	6	917	3	10.03%	0.65%					<69						>0.998		>6.5
N64	6	951	3	10.40%	0.63%					<83						>1.062		>8.5
N65	11	1673	3	18.30%	0.65%					>64						<1.097		>0
N66	5	937	3	10.25%	0.53%					>75						>1.028		>10.9

Figure 8. Negative Theory Consisting of 77 Low-Risk Patterns
The Low-Risk Patterns of **T**

N67	6	933	3	10.21%	0.64%					>79						>1.018		>8.5
N68	5	926	3	10.13%	0.54%					>82							0	>10.9
N69	7	1084	3	11.86%	0.64%										>21.5, <29.5	>0.98		
N70	12	1885	3	20.62%	0.63%										>19.5			>0
N71	6	946	3	10.35%	0.63%										>12	>0.989, <1.028		
N72	6	943	3	10.32%	0.63%										>14.5	>0.998, <1.038		
N73	6	956	3	10.46%	0.62%										>18	>0.989, <1.05		
N74	10	1527	3	16.70%	0.65%										>22.5	>0.862, <1.05		
N75	12	1859	3	20.33%	0.64%										>22.5	>0.924	0	
N76	6	932	3	10.19%	0.64%										>18	>1.062		>7
N77	6	940	3	10.28%	0.63%										>29.5	>0.823		>5

Figure 8. Negative Theory Consisting of 77 Low-Risk Patterns (continued)

8. Prognostic Index

The classification provided by a model assigns a risk category for those patients who satisfy only high-risk or low-risk patterns.

We shall introduce in the this section a classification tool, which refines the classification provided by a model, by assigning a *prognostic index* to each patient, thus making it possible to assign a risk category also to those patients who satisfy both high-risk and low-risk patterns.

Let τ^+ be the number of high-risk patterns in the model **T**, and $\tau^+(x)$ be the number of those high-risk patterns which are satisfied by a particular observation *x*; the indices τ^- and $\tau^-(x)$ are defined in a similar way.

The *prognostic index* $\pi(x)$ of patient *x* is defined as

$$\pi(x)=\frac{\tau^+(x)}{\tau^+}-\frac{\tau^-(x)}{\tau^-}. \tag{10}$$

Clearly, $\pi(x)$ will have a positive value for observations satisfying only high-risk patterns, and a negative value for those satisfying only low-risk patterns. By classifying now an observation *x* as being at high-risk or at low-risk according to the positive or negative sign of $\pi(x)$, we shall refine the classification provided in section 7 (since the classification defined in that section will never be contradicted).

The classification matrix for the prognostic index is defined very similarly to that in section 7, noticing simply that only 3 columns (**HRI, LRI** and **UI**) are needed this time, since the set **HLI** of patients having both positive and negative indices is obviously empty**.**

For the CCF dataset the prognostic index π produces the following classification matrix:

	HRI	**LRI**	**UI**
P	74.68%	23.72%	1.60%
N	18.45%	78.89%	2.66%
Risk	**12.14%**	**1.02%**	
Prev	**20.31%**	**77.07%**	**2.63%**

Figure 9. Classification Matrix for Prognostic Index

Clearly, the high-risk set defined by the prognostic index contains 1920 patients, i.e., 20.31% of the entire population, and has a mortality rate of 12.14%. The low-risk set contains 7286 patients, i.e., 77.07% of the entire population, and has a mortality rate of 1.02%.

The graphs in Figure 10 present the changes in the survival rates of the three index-defined categories of patients (low-risk, unclassified and high-risk) along the nine years of observation, illustrating the sharp differences between them.

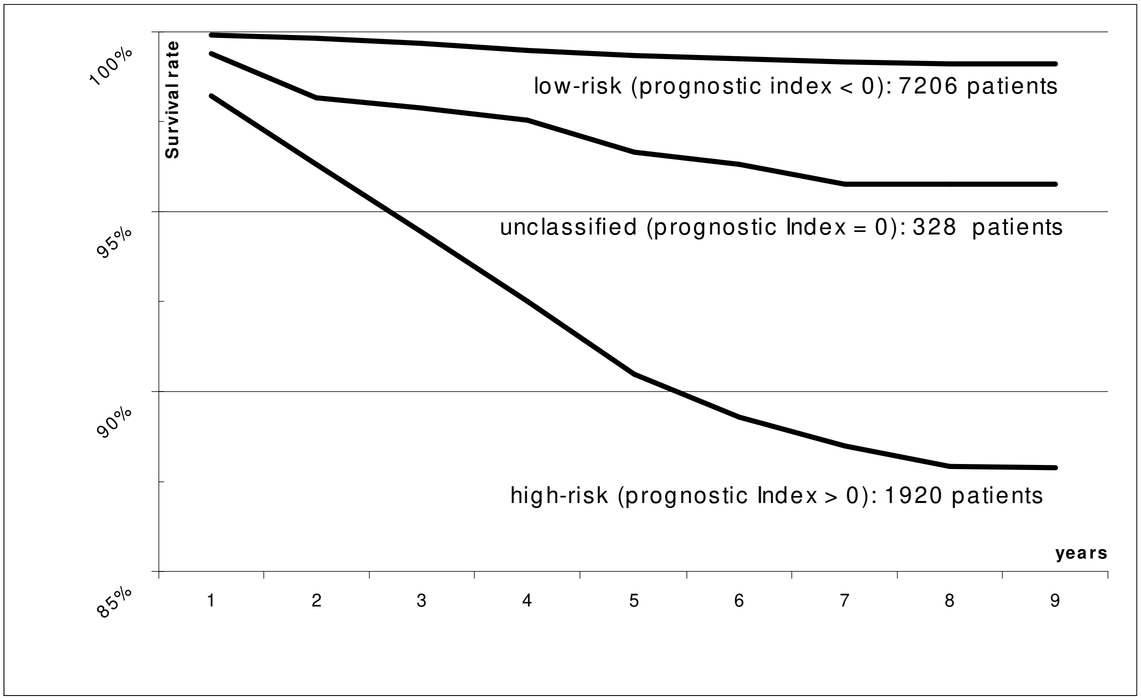

Figure 10. Dynamics of Survival Rates for Low-risk, Unclassified and High-risk Patterns

9. Model validation

In order to evaluate the accuracy of the model we have applied to it the standard 5-folding validation method, i.e., we have considered a random partition of the original dataset into 5 segments, each of them containing about one fifth of the positive and one fifth of the negative observations. Using this partition we have defined 5 LAD problems. In each of them one of the 5 segments of the original dataset was chosen to be used as "testing set" for validating the high-risk and low-risk patterns, as well as the positive and negative models obtained by applying LAD to the "training set", i.e., to the set consisting of the remaining 4 segments of the entire dataset Ω. For each of the 5 problems, using the chosen parameter values (30 cutpoints for each numerical variable, pattern degree at most 3, high-risk threshold 16.5%, low-risk threshold 0.66%, and positive and negative prevalences of at least 10% each) we have first calculated the risk and the relative prevalences of the pandects of high-risk and low-risk patterns of the 5 training sets, and recalculated them on the 5 corresponding testing sets.

The results of these calculations are reported in Figures 11 and 12. It can be seen that the average risk among all the observations satisfying at least one of the high-risk pattern remains above 15% in the testing set, and that the average risk among all those observations which satisfy at least one of the low-risk pattern remains below 0.6%. Also, it can be seen that an average high-risk pattern is satisfied by more than 17% of the positive observations in the testing set, and the average low-risk pattern is satisfied by more than 11% of the negative observations.

	Risk			
Validation problem	High-risk		Low-risk	
	Training	Testing	Training	Testing
1	16.82%	13.04%	0.41%	0.45%
2	16.88%	15.71%	0.41%	0.76%
3	16.86%	15.10%	0.40%	0.51%
4	16.82%	15.86%	0.39%	0.64%
5	16.84%	15.75%	0.40%	0.60%
Average	**16.84%**	**15.09%**	**0.40%**	**0.59%**

Figure 11. Risk Validation of Pandect by 5-Folding

	Relative prevalence			
Validation problem	Positive		Negative	
	Training	Testing	Training	Testing
1	19.54%	17.56%	11.19%	10.76%
2	19.97%	17.06%	11.34%	11.51%
3	18.91%	17.45%	11.09%	11.16%
4	19.30%	17.95%	11.19%	10.81%
5	17.83%	16.07%	11.25%	11.62%
Average	**19.11%**	**17.22%**	**11.21%**	**11.17%**

Figure 12. Relative Prevalence Validation of Pandect by 5-Folding

Figure 13 presents the results of applying the 5-folding validation method to the classification matrices corresponding to the prognostic index. It can be seen that the high-risk sets defined by this index have an average mortality rate of 10.75%, while the low-risk sets have an average mortality rate of only 1.3%.

Validation problem		Training HRI	Training LRI	Training UI	Testing HRI	Testing LRI	Testing UI
1	P	78.00%	20.40%	1.60%	72.58%	24.19%	3.23%
	N	19.44%	78.78%	1.78%	20.35%	78.01%	1.64%
	Risk	**12.06%**	**0.88%**		**10.79%**	**1.04%**	
	Prev	**21.37%**	**76.85%**	**1.77%**	**22.07%**	**76.23%**	**1.69%**
2	P	77.20%	21.60%	1.20%	61.29%	38.71%	0.00%
	N	18.61%	79.63%	1.76%	17.18%	80.96%	1.86%
	Risk	**12.42%**	**0.92%**		**10.80%**	**1.60%**	
	Prev	**20.54%**	**77.71%**	**1.75%**	**18.63%**	**79.57%**	**1.80%**
3	P	78.40%	19.60%	2.00%	75.81%	22.58%	1.61%
	N	19.36%	78.70%	1.94%	20.51%	77.13%	2.35%
	Risk	**12.16%**	**0.84%**		**11.14%**	**0.98%**	
	Prev	**21.31%**	**76.75%**	**1.94%**	**22.34%**	**75.33%**	**2.33%**
4	P	79.60%	20.00%	0.40%	59.68%	35.48%	4.84%
	N	18.55%	79.82%	1.63%	18.93%	79.43%	1.64%
	Risk	**12.79%**	**0.85%**		**9.66%**	**1.49%**	
	Prev	**20.57%**	**77.85%**	**1.59%**	**20.27%**	**77.98%**	**1.75%**
5	P	75.00%	24.19%	0.81%	64.06%	32.81%	3.13%
	N	17.82%	80.88%	1.30%	17.49%	80.93%	1.58%
	Risk	**12.49%**	**1.00%**		**11.36%**	**1.40%**	
	Prev	**19.71%**	**79.01%**	**1.28%**	**19.02%**	**79.34%**	**1.64%**

	Average HRI	Average LRI	Average UI	Average HRI	Average LRI	Average UI
P	77.64%	21.16%	1.20%	66.68%	30.76%	2.56%
N	18.76%	79.56%	1.68%	18.89%	79.29%	1.82%
Risk	**12.38%**	**0.90%**		**10.75%**	**1.30%**	
Prev	**20.70%**	**77.63%**	**1.67%**	**20.47%**	**77.69%**	**1.84%**

Figure 13. Prognostic Index Validation by 5-Folding -- Classification Matrices

10. Conclusions

Using Logical Analysis of Data on a population of patients under exercise stress testing, we have shown that it is possible to reliably identify a small subset of patients who are at relatively high risk for death while simultaneously identifying a large population of patients who are at very low risk.

A characteristic feature of this dataset consisted in the large disproportion between the number of the "positive" and of "negative" observations. This disproportion is due to the nature of the two classes of observations, which represented respectively the groups of patients who died or who survived during the observation period. While this disproportion is entirely reasonable in many medical (and some other) datasets, the methodological implications of it have not been previously examined in any other real-life application of LAD. From a purely methodological point of view this study led to an extension of the scope of LAD to the case of non-separable datasets.

A remarkable feature of the classification provided by LAD consists in the fact that the 20% segment of the population identified as being at high-risk included 3/4 of all those who died during the observation period; the mortality rate of this segment was 4 times higher than the average mortality rate. In contrast the 77% segment of segment of the population identified as being at low-risk had a mortality rate of less than 1/3 of the average mortality rate. Finally, the size of the unclassified segment of the population was of only 2.63%.

Most medical literature on risk stratification has focused on specific predictors of risk, with relatively less emphasis on interaction, that is, on ways in which predictor variables affect each other's impacts on risk of an adverse outcome. While careful multivariable modeling makes it possible to examine two-way interactions, LAD, by its very nature, makes it possible to examine thousands of possible interactions with high degrees of complexity. It can be expected that the interactions revealed through LAD may stimulate research for a better understanding of the related cause-effect relationships.

Among the useful features of LAD, we mention the possibilities it offers for explaining or justifying -- on the basis of the patterns triggered by the values of the variables corresponding to a particular patient -- the decision to classify the patient into a high-risk or a low-risk class, as well as decisions concerning the choice of particular alternatives for his or her treatment.

References

[1] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, Logical Analysis of Numerical Data, *Mathematical Programming*, 79, 1997, pp. 163-190

[2] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik, An Implementation of Logical Analysis of Data, *IEEE Transactions on Knowledge and Data Engineering*, 12, No. 2, 2000, 292-306

[3] R.M. Califf, P.W. Armstrong, J.R. Carver, R.B. D'Agostino, W.E. Strauss. 27th Bethesda Conference: Matching the Intensity of Risk Factor Management with the Hazard for Coronary Disease Events. Task Force 5. Stratification of Patients into High, Medium and Low Risk Subgroups for Purposes of Risk Factor Management. *J Am Coll Cardiol* 27(5), 1996, 1007-19

[4] C.R. Cole, E.H. Blackstone, F.J. Pashkow, C.E. Snader, M.S. Lauer. Heart-Rate Recovery Immediately after Exercise as a Predictor of Mortality. *N Engl J Med* 341(18), 1999, 1351-57

[5] D. Cox. Regression Models and Life Tables (with discussion). *J R Stat Soc B* 34, 1972, 187-220.

[6] Y. Crama, P.L. Hammer, T. Ibaraki, Cause-Effect Relationships and Partially Defined Boolean Functions, *Annals of Operations Research.* 16 (1988) 299-326

[7] O. Ekin, P.L. Hammer, A. Kogan, Convexity and Logical Analysis of Data, *Theoretical Computer Science*, 244, 2000, 95-116

[8] P.L. Hammer, Y. Liu, B. Simeone, S. Szedmak, Saturated Systems of Homogeneous Boxes and the Logical Analysis of Numerical Data, *Proceeding 24th Annual Conference, German*

Gesellschaft fur Klassifikation, Passau, Germany, 2000, and *RUTCOR Research Report RRR 19-2000,* RUTCOR, Rutgers University, Piscataway NJ 08854

[9] P.L. Hammer, Partially Defined Boolean Functions and Cause-Effect Relationships, *International Conference on Multi-Attribute Decision Making Via OR-Based Expert Systems*, University of Passau, Passau, Germany, 1986.

[10] D. Hosmer, S. Lemeshow. *Applied Logistic Regression*. New York: Wiley; 1989.

[11] J.M. Kwok, T.D. Miller, T.F. Christian, D.O. Hodge, R.J. Gibbons. Prognostic Value of a Treadmill Exercise Score in Symptomatic Patients with Nonspecific ST-T Abnormalities on Resting ECG. *J. Amer. Medical Assoc.* 282(11), 1999, 1047-53.

[12] M.S. Lauer, G.S. Francis, P.M. Okin, F.J. Pashkow, C.E. Snader, T.H. Marwick. Impaired Chronotropic Response to Exercise Stress Testing as a Predictor of Mortality. *J. Amer. Medical Assoc.* 1999(6), 1999, 524-29

[13] D.B. Mark, M.A. Hlatky, F.E. Harrell, Jr., K.L. Lee, R.M. Califf, D.B. Pryor. Exercise Treadmill Score for Predicting Prognosis in Coronary Artery Disease. *Ann Intern Med* 106(6), 1987, 793-800

[14] D.B. Mark, L. Shaw, F.E. Harrell, Jr., M.A. Hlatky, K.L. Lee, J.R. Bengtson, et al. Prognostic Value of a Treadmill Exercise Score in Outpatients with Suspected Coronary Artery Disease [see comments]. *N Engl J Med* 325(12), 1991, 849-53

[15] T.H. Marwick, R. Mehta, K. Arheart, M.S. Lauer. Use of Exercise Echocardiography for Prognostic Evaluation of Patients with Known or Suspected Coronary Artery Disease. *J Am Coll Cardiol* 30(1), 1997, 83-90

[16] E.O. Nishime, C.R. Cole, E.H. Blackstone, F.J. Pashkow, M.S. Lauer. Heart Rate Recovery and Treadmill Exercise Score as Predictors of Mortality in Patients Referred for Exercise ECG. *J. Amer. Medical Assoc.* 284(11), 2000, 1392-8

[17] G. Rose. *The Strategy of Preventive Medicine.* New York: Oxford University Press; 1992.

[18] D.A. Weiner, T.J. Ryan, Parsons L, L.D. Fisher, B.R. Chaitman, L.T. Sheffield, et al. Long-Term Prognostic Value of Exercise Testing in Men and Women from the Coronary Artery Surgery Study (CASS) Registry. *Am J Cardiol* 75(14), 1995, 865-70.