RUTCOR
RESEARCH
REPORT

Logical Analysis of Data: From Combinatorial Optimization to Medical Applications

Peter L. Hammer[a] Tibérius Bonates[b]

RRR 10 - 2005, February 2005

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]RUTCOR, Rutgers University, Piscataway, NJ (hammer@rutcor.rutgers.edu).
[b]RUTCOR, Rutgers University, Piscataway, NJ (tbonates@rutcor.rutgers.edu).

Rutcor Research Report
RRR 10 - 2005, February 2005

Logical Analysis of Data: From Combinatorial Optimization to Medical Applications

Peter L. Hammer Tibérius Bonates

Abstract. After reviewing the basic concept of the Logical Analysis of Data (LAD), the paper presents a series of discrete optimization models associated to the implementation of various components of the general methodology of LAD, and concludes with an outline of applications of LAD to medical problems. The combinatorial optimization models described in the paper represent variations on the general theme of set covering, including some with nonlinear objective functions. The medical applications described include the development of diagnostic and prognostic systems in cancer research and pulmonology, risk assessment among cardiac patients, and the design of biomaterials.

Acknowledgements: The first author gratefully acknowledges the partial support of the National Science Foundation (grant NSF-IIS-0312953), and the National Institutes of Health (award numbers HL-072771-01 and NIH-002748-001). The second author gratefully acknowledges the partial support of a DIMACS Graduate Student Award. We also acknowledge the assistance provided by Dash Optimization by allowing the use of its linear and integer programming solver Xpress-MP within its Academic Partnership Program.

1 Introduction

1.1 Medical Data Analysis

A large number of typical data analysis problems appearing in medicine and in numerous other areas can be formulated in the following way. A "dataset" consisting of two disjoint sets Ω^+ and Ω^- of n-dimensional real vectors is given. Typically each of the vectors appearing in the dataset correspond to a patient, the vectors in Ω^+ corresponding to patients having a specific medical condition (e.g. pneumonia), while those in Ω^- (the "controls" in medical language) do not have that condition. The components of the vectors, called "attributes", or "features", or sometimes "variables", can represent the results of certain measurements, tests, the expression levels of genes or proteins in the blood of the patients, or can simply indicate the presence or absence of certain symptoms (in which case they are usually expressed as zeros or ones).

Diagnosis is one of the typical questions arising in the analysis of such data. In this case the problem is simply to extract information (i.e. to "learn") from a given dataset in order to recognize whether a "new" patient, i.e. an n-vector not contained in the dataset, does or does not have the specific condition under analysis.

Prognosis is a similar problem. In this case it is assumed that the vectors in the dataset are known to have or not to have developed a particular medical condition (e.g. a cardiac event, cancerous metastases, etc) within a well-defined time period (typically 5 or 10 years). In this case again the problem is to learn enough from the given data to predict whether a new patient is prone to develop within that time period the condition under analysis.

Diagnosis and prognosis are two special cases of what is called *classification* in data analysis, and considered by many to be its central problem. The basic idea of classification is to "learn", i.e. to extract enough information from the dataset to be able to recognize the positive or negative nature of a new point.

The identification of individualized therapies – on the basis of data analysis and mathematical/computational diagnostic and prognostic systems – is a major challenge of this new area of "medical bioinformatics" [11].

1.2 Illustrative Examples and Real Life Applications

In the last section we shall present some of the significant medical problems to which LAD has been actually applied in a series of collaborative studies with biomedical researchers from Cancer Institute of New Jersey, Cleveland Clinic Foundation, Food and Drug Administration (FDA), Hôpital Avicenne (Paris), National Institutes of Health (NIH), New Jersey Center for Biomaterials, Robert Wood Johnson Medical School, University of Grenoble, various centers and departments of Rutgers University, etc. Additional ongoing collaborative efforts for applying LAD to the analysis of medical data involve researchers from the NIH Clinical Center for Radiology and Imaging Sciences, the NIH Clinical Proteomic Applications Center, the Nephrology Division of Mount Sinai School of Medicine (New York), Semelweiss Medical University (Budapest), Eötvös Loránd Science University (Budapest), etc.

These studies include the development of diagnostic and prognostic systems for ovarian cancer, breast cancer and lymphoma, the differentiation of various types of idiopathic pneumonia, risk assessment among cardiac patients, the design of biomaterials, etc. The datasets used in these examples are very considerably in character and size. The various datasets include clinical, genomic, proteomic, computer tomography and polymeric data, the sets of variables considered ranging from a few dozens to 25,000, and the number of patients from less than a hundred to 10,000.

Since these datasets are of considerable size, in the section outlining the basic ideas of the methodology of the Logical Analysis of Data (LAD) we shall use four considerably smaller datasets which are available on the Web, and which have been frequently used in numerous studies of medical data analysis. The datasets [1] Heart Disease (HD), Pima Indians' Diabetes (PID), Breast Cancer Wisconsin (BCW), and BUPA Liver Disease (BLD) will serve as such illustrative examples.

The basic parameters of these four datasets are the following:

	# of Observations		# of Attributes	
Dataset	Positive	Negative	Numerical	Binary
Breast Cancer Wisconsin (BCW)	236	213	9	0
Heart Disease (HD)	137	160	10	3
Pima Indian Diabetes (PID)	130	262	8	0
Bupa Liver Disorders (BLD)	200	145	6	0

Table 1: Parameters of datasets.

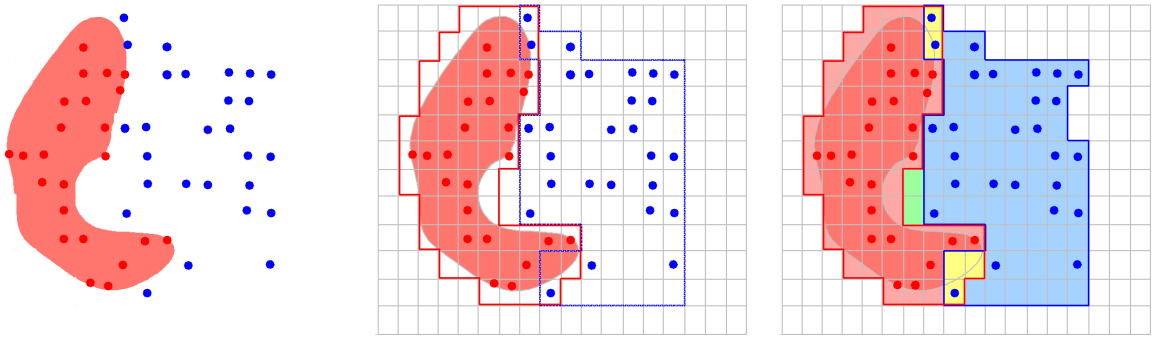
Remark: It is known from the literature that BCW is a "clean" dataset on which many data analysis methods provide highly accurate diagnostic models. HD is somewhat less clean than BCW, but still reasonably predictable. On the other hand, it is known that for the problems PID, and especially for BLD it is very hard to find accurate computational models.

1.3 Principles of LAD

The Logical Analysis of Data (LAD) is a combinatorics and optimization based data analysis method [14, 16, 19]. While LAD has been applied to numerous disciplines, e.g. economics and business, seismology, oil exploration, etc. (see [14], [18], [21], [22], etc.), in this paper we shall deal only with its applications to medicine and related disciplines.

The basic idea of LAD is to combine a differentiation/integration approach of a subspace of R^n containing the given positive observations and negative observations, and the "new" (i.e. not yet seen) ones. In the "differentiation" step a family of small subsets having strong positive or negative characteristics is identified. In the "integration" step unions of certain subsets of such positive (respectively, negative) subsets are proposed as approximations of the areas of R^n containing the positive (respectively, negative) "new" or "old" observations. Concretely, the basic components of LAD are the following:

[1] See http://www.ics.uci.edu/∼mlearn/MLRepository.html.



(A) (B) (C)

Figure 1: Naive representation of dataset and (A) Nature's positive (red) and negative (blue) zones; (B) LAD's approximation of the two zones; (C) Classification based on LAD's approximations (Red: +, blue: −, yellow: + or − depending on discriminant, green: unclassified).

(a) In order to eliminate redundant variables in the original dataset we extract from it a (usually minimal) subset S, capable of distinguishing the positive observations from the negative ones. In the subsequent steps we work mostly with the projections Ω_S^+ and Ω_S^- of Ω^+ and Ω^-, respectively, on this subset of variables. While this so-called "feature extraction" step is present in most data analysis methods, the specific way in which it is applied within the LAD methodology emphasizes the interaction of variables, and the importance of retaining not only those which can influence individually the positive or negative nature of observations, but also those whose "collective" or "combinatorial" effect is significant.

(b) We cover Ω_S^+ with a family of (possibly overlapping) homogeneous subsets of the reduced real space, each of these subsets having a significant intersection with Ω_S^+, but being disjoint from Ω_S^-. The only subsets considered by LAD are intervals of Ω_S^+ with parallel faces to the axis; these intervals are called "positive patterns".

(b') A similar construction is applied to Ω_S^- for finding "negative patterns".

(c) A subset of the positive (respectively, negative) patterns, the union of which covers every observation in Ω_S^+ (respectively, Ω_S^-) is identified. The collection of these two subsets of intervals is called a "model".

(d) A classification method is developed which defines the positive or negative character of any observation covered by the union of the two subsets of intervals of the model, leaving as "unclassified" only those observations which are not covered by this union.

(e) One of the typical validation methods is applied to verify the accuracy of the resulting classification system.

Figure 1 illustrates part of the steps described above. Details of the LAD components will be briefly outlined in section 2. It is interesting to know that many of the mathematical problems needed in handling these components require the solution of different variations of the well-known set covering problem.

1.4 Why LAD?

Although the origins of LAD go back to 1986, its practical applications to medical problems was only started in 2002-2003, with the publication [9, 26] of the conclusions of a collaborative study with medical researchers at the Cleveland Clinic Foundation on risk assessment among cardiac patients. In the last two decades numerous other data analysis methods have been elaborated and very successfully applied to the analysis of medical data. Among other frequently applied methods, we mention Neural Networks, Nearest Neighbors, Decision Trees, Support Vector Machines, etc. The reasons for which LAD has been applied to the problems described in this paper are multiple, and we shall only mention some.

The most specific feature of LAD which has substantial implications on its applicability is that instead of just asking the question whether a new observation is positive or negative, it tries to approximate the subspace of R^n containing the positive and that containing the negative observations. Among the most important practical consequences of this approach we mention the possibility of providing explanations for each diagnostic or prognostic conclusion of LAD, the possibility of identifying new classes of observations and of analyzing the role and nature (e.g. blocking or promoting) of features [6], the possibility of developing individualized therapies for patients [11], the development of catalogs of research hypothesis for medical researchers, etc.

Dataset	SVM	C4.5	Nearest Neighbors	Neural Networks	LAD [2]
BCW	95.3%	93.5%	94.1%	93.7%	95.0%
HD	83.4%	82.7%	75.7%	80.5%	83.4%
PID	77.9%	77.7%	72.6%	74.5%	77.2%
BLD	70.8%	66.2%	62.2%	69.1%	74.9%

Table 2: Diagnosis accuracy of some frequently used data analysis methods.

(a) The first question which has to be answered about the choice of methodologies concerns the accuracy of diagnosis and prognosis, i.e. the proportion of correctly classified points in randomly generated sets of "new" points.

In order to present some information about the diagnostic accuracy of various data analysis methods, we present in Table 2 the results obtained by using the methods of *Support Vector Machines (SVM)*, *Decision Trees (C4.5)*, *Nearest Neighbors*, *Neural Networks (Multilayer Perceptron)*, and the *Logical Analysis of Data (LAD)*. For the first four methods we

[2] Using approximately maximum patterns, obtained by solving L_2 best linear approximation of polynomial set covering problem.

used the software included in the Weka package [30], while for LAD we used RUTCOR's software. All results have been validated through 20 10-folding experiments.

It can be seen that the accuracy of LAD is very comparable with that of the best methods used in data analysis, providing usually closely resembling results with the four most frequently used methods. The value of LAD was reconfirmed and considerably strengthened in the real life medical applications to be described in section 3, in which we have obtained accuracies substantially exceeding those using other methods, as reported by other researchers in the literature.

(b) One of the specific features of Logical Analysis of Data (LAD) is that, in contrast to different "black box"-type methods, which provide the classification of new points without any explanations, LAD provides for each classification an justification of the reasons for which a patient is viewed by LAD as being positive or negative. Each explanation of the specific reasons for which a particular patient is classified by LAD as, say positive, has two components: (i) the LAD classification presents the list of positive patterns displayed by the patient, which is, of course, displayed by large proportions of the positive cases in the dataset, but by none of the negative ones, and (ii) evidence for the fact that the patient in question does not satisfy the defining conditions of any of the negative patterns in the model. The utility of such explanations in arriving to a medical decision is obvious for doctors, patients, hospitals, insurance companies, the pharmaceutical industry, and various government agencies involved in health care.

(c) Although diagnosis and prognosis are essential components of the analysis of medical data, numerous other information can be learned from datasets [6]. Among the most interesting conclusions reported in the literature we mention the discovery of "biomarkers" (i.e. highly influential variables), "combinatorial biomarkers" (i.e. highly influential combinations of pairs or triplets of variables), "blockers" or "promoters" (i.e. variables favoring or inhibiting a certain medical condition), new classes of patients and new classes of variables having highly similar properties, etc.

(d) Another interesting result of the mathematical analysis of datasets is the formal identification of numerous specific patterns distinguishing patients from controls. The need to understand and explain the biomedical mechanisms at the basis of a pattern's function associate in a natural way research hypotheses to each of them and can be of major assistance to medical researchers.

(e) The global perspective offered by LAD on the subspace of positive and negative observations allows non-usual applications ranging from the design of biomaterials or drugs to the development of individualized therapies which takes into account the specifics of each patient. As examples of these possibilities we mention the predictive model developed in [2] concerning the correlation of the chemical composition of a biomaterial with the biological response of cells, and the method proposed in [11] for identifying breast cancer therapies consisting in the inhibition of specific small, patient-dependent, subsets of kinases.

2 Logical Analysis of Data

We shall present below several basic techniques of LAD applicable to binary datasets, including techniques for eliminating redundant attributes ("support set selection"), identifying large homogeneous subcubes of the unit cube (i.e. subcubes containing only, say positive, observations), constructing "models" (i.e. two families of such subcubes such that Ω^+ is contained in the union of one of them, and Ω^- is contained in the union of the other). Before presenting these techniques we shall first show in subsection 2.1 a natural way of transforming a numerical problem to a binary one.

2.1 Binarization

The Logical Analysis of Data was originally developed for the analysis of datasets whose attributes take only binary (0-1) values [19, 16, 14]. Since it turned out later that most of the real-life applications include attributes taking real values, a "binarization" method was proposed in [13].

The basic idea of binarization is very simple. It consists in the introduction of several binary attributes associated to each of the numerical attributes; each of these binary attributes is supposed to take the value 1 (respectively, 0) if the numerical attribute to which it is associated takes values above (respectively, below) a certain threshold. Obviously the computational problem associated to binarization is to find a minimum number of such threshold values (cutpoints) which preserve the essential information contained in the dataset, i.e. the disjointness of the sets of (binarized) positive and negative observations. For illustration, Figure 2 shows two binarizations of the same dataset, using respectively 8 and 4 threshold values.

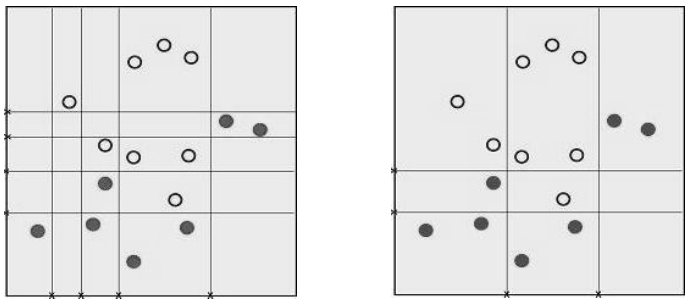


Figure 2: Arbitrary and optimal binarization of same dataset.

It has been shown in [13] that the solution of this minimization problem is NP-hard, even in apparently simple cases (e.g. for 2-dimensional problems). We shall present below a model for this problem for which various heuristics have been developed and are currently used in the solution of practical problems. First, let us remark that if we order the patients according to non-decreasing values of a particular attribute X, and if this attribute takes the values

$X_1 \leq X_2 \leq \cdots \leq X_{|\Omega|}$, then in order to maintain the disjointness of the sets of binarized images of the points in Ω_S^+ and Ω_S^-, (i) it is sufficient to consider only those threshold values t which belong to intervals (X_i, X_{i+1}), where X_i and X_{i+1} are the X-coordinate values of two patients having different "signs" (one being positive and the other one negative), (ii) it is sufficient to consider at most one threshold value (say, $\frac{X_i+X_{i+1}}{2}$) in each of the above described intervals (X_i, X_{i+1}).

Based on these remarks we can form a collection consisting of one cutpoint for each relevant interval corresponding to each numerical attribute. For a typical problem in medical applications (especially in proteomics and genomics), the number of numerical attributes is in the tens of thousands, and the number of such potential cutpoints for each of them is in the hundreds. If we associate a 0-1 variable to each potential cutpoint, in order to solve the minimization problem of the number of cutpoints needed for binarization, we have to solve a combinatorial optimization problem with millions of binary variables. Because of obvious computation limitations, the consideration of such a problem must be preceded by the use of simple heuristic techniques to eliminate from consideration the vast majority of potential cutpoints.

A number of simple statistical, information-theoretical, and combinatorial pre-processing techniques have been developed [7] for the solution of this problem. The simplest statistical technique consists in evaluating the correlation of each of the binary vectors corresponding to the cutpoints with the binary vector of "outcomes" (which are usually defined as being 1 for the positive vectors and 0 for negative ones); the binary features whose correlation with the outcome is sufficiently low are simply eliminated from consideration. The simplest combinatorial technique for "feature elimination" associates simply to each binary vector the number of those pairs consisting of a positive and a negative observation which differ in that binary feature; binary vectors which distinguish in this way small numbers of pairs of positive and negative observations are eliminated from consideration. Usually several such simple heuristics are applied iteratively until the number of features remaining is reduced to a manageable size (hundreds or at most thousands).

After all these simplifications, the remaining binary problem is treated with the usual LAD techniques, including – among others – feature elimination through the identification of a minimum size support set. This and several other typical LAD techniques will be described below.

In order to illustrate the binarization of medical datasets, let us consider the examples presented in Table 1. A very simple binarization procedure introduces for each numerical variable x a fixed number (in our example, 3) of cutpoints ψ_1, ψ_2, ψ_3. The cutpoints are introduced in such a way that the 4 subsets of observations for which $x \leq \psi_1$, $\psi_1 < x \leq \psi_2$, $\psi_2 < x \leq \psi_3$ and $\psi_3 \leq x$ should have approximately the same cardinality. If in the resulting binarized problem the images of a positive and a negative observation coincide then the chosen binarization is infeasible, and the number of cutpoints used for some of the variables has to be increased. This process is continued until the images of Ω^+ and Ω^- in the binarized space become disjoint. We present in Table 3 the number of variables appearing in the binarized versions of the datasets BCW, PID, HD and BLD binarized as above.

Dataset	# of Attributes		# of Binarized Attributes
	Numerical	Binary	
BCW	9	0	60
HD	10	3	63
PID	8	0	64
BLD	6	0	43

Table 3: Binarization of datasets.

2.2 Support Sets and Biomarkers

Whether a binary dataset is obtained by binarizing a numerical dataset or is generated naturally, it is very likely to contain a number of redundant attributes. We shall describe below a simple combinatorial optimization model for eliminating sets of redundant variables.

A set S of attributes is called a *support set* if the projection Ω^+_S of Ω^+ on S is disjoint from the projection Ω^-_S of Ω^- on S. The complete set $N = 1, \ldots, n$ is a support set, since we have assumed Ω^+ and Ω^- to be disjoint. A support set is called minimal or irredundant if by eliminating any one of its variables the remaining dataset will contain a positive observation and a negative observation which coincide.

In order to identify a minimal support set, let us associate now to every attribute x_i, $i = 1, \ldots, t$ in the binary dataset, a new binary variable y_i equal to 1 if variable x_i belongs to the support set, and equal to 0 if it does not. Let further $U = (u_1, u_2, \ldots, u_t)$ be the binary vector associated to a positive patient and $V = (v_1, v_2, \ldots, v_t)$ be the binary vector associated to a negative patient. Let us further associate to this pair consisting of a positive and a negative observation the vector $(w_1(U,V), \ldots, w_t(U,V))$, where $w_i(U,V) = u_i \oplus v_i (\text{mod } 2)$, i.e. $w_i(U,V) = 1$ if $u_i \neq v_i$, and $w_i(U,V) = 0$ otherwise. Obviously, the condition that the projections of binarized images Ω^+_{BS} and Ω^-_{BS} of Ω^+_S and Ω^-_S on S should be disjoint, is equivalent to requiring that for any $U \in \Omega^+_{BS}$ and $V \in \Omega^-_{BS}$,

$$\sum w_i(U,V) y_i \geq 1. \tag{1}$$

The minimum size support sets can be simply obtained by solving the set covering problem

$$\begin{aligned} min \quad & \sum_{j=1}^{t} y_j \\ s.\,t. \quad & \sum w_i(U,V) y_i \geq 1, \qquad \text{for all } U \in \Omega^+_{BS} \text{ and } V \in \Omega^-_{BS} \\ & y \in \{0,1\}^t. \end{aligned}$$

We present in Table 4 the cardinalities of the minimum support sets of the datasets BCW, HD, BL, PID binarized by solving this set covering problem.

Dataset	# of Attributes		# of Binarized Attributes	Size of Minimum Support Set
	Numerical	Binary		
BCW	9	0	60	11
HD	10	3	63	14
PID	8	0	64	16
BLD	6	0	43	18

Table 4: Size of minimum support sets.

Two important variations of this problem play a special role in finding support sets. First, because of small variations in the measurements obtained by using different instruments in different laboratories, it is important that a support set should be as "stable" or "robust" as possible. In order to achieve this goal, in many applications we replace the usual set covering constraint (1) by the stronger requirement

$$\sum w_i(U,V)y_i \geq d, \tag{2}$$

where d is a positive integer chosen so as to assure the robustness of the outcome. We shall call this problem the *d-covering* or *multiple covering* problem. In many applications d is taken equal to 3, 5, 10 or even 20, depending on the number of variables in the original formulation of the problem.

A second variation which is frequently important to be added to the formulation of the above set covering problem translates the requirement that certain pairs of binary variables cannot be simultaneously present in a support set. For example, in the case of binarized problems it is important that the cutpoints which define any pair of binarized variables associated to the same numerical variable should be at a sufficiently large distance to compensate for possible imprecisions in measurement. Therefore, if c' and c'' are two cutpoints for the binarization of the same numerical variable, and if y' and y'' are the decision variables associated to the binarized variables x' and x'' corresponding to those cutpoints, then it is natural to require that for sufficiently small δ values,

$$\text{if } |c' - c''| \leq \delta \text{ then } y' + y'' \leq 1.$$

A frequently occurring concept in the biomedical literature is that of "biomarkers", i.e. variables with a great influence on the outcome of a phenomenon. It is not unreasonable to designate the variables appearing in irredundant support sets as biomarkers. The selection of these variables takes into account not only their individual interaction with the outcome, but also the interactive role of groups of these selected variables. The next subsection will deal with the identification of minimum interactive sets of variables (*patterns* or *combinatorial biomarkers*) capable of indicating the positive or negative nature of an observation.

2.3 Patterns and Maximum Patterns

A basic concept in the analysis of data is that of a *pattern*. A *positive pattern* is simply a subcube of the unit cube which intersects Ω^+_{BS} and is disjoint from Ω^-_{BS}. *Negative patterns* have a similar definition.

Because patterns play a central role in LAD, various types of patterns (e.g. prime, spanned, maximum) have been studied, algorithms have been developed for their enumeration [10, 3, 12, 17], and their relative efficiency has been analyzed [20]. We shall present in this paper one of these pattern types which proved to be particularly useful for LAD.

A positive ω-pattern for $\omega \in \{0,1\}^t$, is a pattern *covering* (i.e., containing) ω. A *maximum positive ω-pattern* P is a positive ω-pattern, the *coverage* of which (i.e., the cardinality of $|P \cap \Omega^+_{BS}|$) is maximum. A maximum negative ω-pattern is defined in a similar way.

Because of the usefulness of this concept in data analysis we shall describe below a combinatorial optimization formulation of the problem of finding a maximum ω-pattern for each observation ω in the dataset. For this purpose, let us first associate to the binary vector $\omega = (\omega_1, \ldots, \omega_t) \in \Omega^+_{BS}$ an "elementary" conjunction C, i.e. a product of some complemented and some uncomplemented variables. Let us define the binary decision variables y_i $(i = 1, \ldots, t)$ in the following way: (i) if $\omega_i = y_i = 1$ then x_i is included in C, (ii) if $\omega_i = 0, y_i = 1$ then $\overline{x}_i$ is included in C, (iii) if $y_i = 0$ then both x_i and $\overline{x_i}$ are absent from C. For example, the decision variables $(0,0,1,1,0)$ associate the elementary conjunction $\overline{x}_3 x_4$ to the observation $\omega = (1,0,0,1,1)$.

The condition that the ω-pattern should not cover any observation from Ω^-_{BS} requires that for every point ρ of Ω^-_{BS}, the variable y_j should take the value 1 for at least one of those j's for which $\rho_j \neq \omega_j$, i.e.

$$\sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^{t} y_j \geq 1, \quad \text{for every } \rho \in \Omega^-_{BS}.$$

On the other hand, a positive observation σ will be covered by the ω-pattern if and only if $y_j = 0$, for all those indices j for which $\sigma_j \neq \omega_j$. Therefore, the number of positive points covered by the ω-pattern will be given by

$$\sum_{\sigma \in \Omega^+_{BS}} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^{t} \overline{y}_j.$$

In conclusion, the maximum ω-pattern problem can be formulated as the following *polynomial*

set covering problem

$$\begin{aligned} max \quad & \sum_{\sigma \in \Omega^+_{BS}} \prod_{\substack{j=1 \\ \sigma_j \neq \omega_j}}^{t} \overline{y}_j \\ s.\,t. \quad & \sum_{\substack{j=1 \\ \rho_j \neq \omega_j}}^{t} y_j \geq 1, \quad \text{for every } \rho \in \Omega^-_{BS} \\ & y_j \in \{0,1\}, \quad \text{for every } j = 1, \ldots, t, \end{aligned}$$

or a strengthened version of it, similar to (2), where in the right hand sides of the set covering inequalities we replace 1 by a positive integer d.

One way of attacking this problem is to linearize it, by introducing a new 0-1 variable for each term of the objective function, and an additional set of linear constraints to guarantee the equality between the values of terms and those of the associated artificial variables. The drawback of this approach is that the large number of artificial variables increases substantially the size, and therefore the difficulties of solving the resulting integer program.

We have proposed in [12] the use of an approximate linearized version of the problem by associating to the objective function the best L_2-norm linear approximation of it, as defined in [23]. It was shown in [23] that the best linear approximation in the norm L_2 of an elementary conjunction $\prod_{j \in T} z_j$ of binary variables x_j is given by the formula

$$-\frac{|T|-1}{2^{|T|}} + \frac{1}{2^{|T|-1}} \sum_{j \in T} z_j.$$

Applying this formula to each term of the objective function of the polynomial set covering problem described above, associates to it a regular set covering problem.

Dataset	Average Prevalences of Maximum Patterns		
	Maximum Patterns	Approx. Max. Patterns [3]	Approx./Exact
BCW	54.2%	46.7%	84.9%
HD	25.6%	23.8%	93.0%
PID	16.3%	14.1%	85.6%
BLD	10.3%	8.8%	85.3%

Table 5: Prevalences of maximum patterns and approximate maximum patterns.

In order to measure the quality of the approximate solutions obtained using the linear approximation of the objective function, we shall use the concept of *prevalence* of a pattern. The *prevalence of a positive (negative) pattern* is simply the proportion of positive (negative) points in the dataset which are covered by that pattern. Table 5 shows the average

[3] Using L_2-best linear approximation of polynomial optimization model.

prevalences of the maximum patterns in the 4 datasets used for illustration. The table also shows the average prevalences of the approximately maximum patterns obtained by solving the linear approximation of the polynomial set covering problem described above. It can be seen that the approximately maximum patterns are of high quality, since their prevalences represent almost 90% of those of the maximum patterns.

2.4 Models: Diagnosis and Prognosis

Diagnosis and prognosis are frequently viewed as the most important applications of data analysis in medicine. Clearly, the knowledge of patterns can be used for deriving conclusions concerning diagnosis and prognosis of yet unseen patients. Indeed, the fact that the measurements of a new patient satisfy the defining conditions of a, say positive, pattern, and at the same time they do not satisfy the defining conditions of any known negative pattern, gives an indication that the patient belongs to the positive group. At the same time it is clear that each pattern, taken in isolation, can be viewed as representing a sufficient condition for a new observation to be positive or negative. Continuing this reasoning, one could expect that the knowledge of the family of all patterns could completely define the positive or negative nature of a new observation, and therefore one can expect to be able to derive from this set necessary and sufficient conditions to specify the sign of a new patient. Because of the exponential number of all possible patterns, this reasoning can only be applied by using a crude approximation of the family of patterns used. In previous successful practical applications we have used for this purpose different kinds of pattern families (prime, spanned, maximum). Since in a model it is always sufficient to consider for each given observation only one of the maximum patterns covering it, the size of the family of all maximum patterns to be involved in a model can be limited to $|\Omega^+ \cup \Omega^-|$. This remark motivates the use of models containing only maximum patterns. We shall show below how to further restrict the number of patterns used in such a model.

Let $\mathcal{M}$ be the family of all maximum patterns in a binary dataset, and let $M_1^+, \ldots, M_p^+$ and $M_1^-, \ldots, M_q^-$ be respectively the sets of positive and negative maximum patterns in $\mathcal{M}$. The following simple classification rules have been seen in numerous examples to provide a diagnostic/prognostic system the accuracy of which compares favorably with that of other bioinformatic systems: (i) if an observation satisfies the defining conditions of some positive patterns, but does not satisfy the defining conditions of any of the negative patterns, then the observation is classified as positive; (ii) if an observation satisfies the defining conditions of some negative pattern, but does not satisfy the defining conditions of any one of the positive patterns, then the observation is classified as negative; (iii) if the observation satisfies the defining conditions of p' of the p positive patterns in $\mathcal{M}$, as well as those of q' of the q negative patterns in $\mathcal{M}$, then the observation is predicted to have the sign of $(\frac{p'}{p}) - (\frac{q'}{q})$; (iv) in the (highly unlikely) event that a new observation does not satisfy the defining conditions of any of the positive or negative patterns in $\mathcal{M}$, the observation is left *unclassified*.

The above defined classification rules have been verified experimentally to hold on numerous datasets, and are considered by many as a useful instrument of biomedical informatics.

It is natural, however, to ask whether there is any redundancy in the number of patterns used for classification. One way of determining a *model*, i.e. a subset of patterns capable of providing classifications for the same set of points which can be classified by the complete system $\mathcal{M}$, is based again on the solution of set covering type problems. In order to formulate these problems we shall associate 0-1 variables r_j $(j = 1, \ldots, p)$ and s_h $(h = 1, \ldots, q)$ to each of the positive and negative patterns in $\mathcal{M}$, with the interpretation that a pattern is included in the model to be constructed, if and only if the corresponding variable takes the value 1.

Obviously, a necessary and sufficient condition for a family of patterns to represent a model is that each point in Ω^+_{BS} should be covered by at least one of the patterns in the set $\{M^+_1, \ldots, M^+_p\}$, and each point in Ω^-_{BS} should be covered by at least one pattern in the set $\{M^-_1, \ldots, M^-_q\}$. If we denote by $\omega \in M$, for a positive or negative pattern M, the fact that a given observation ω is covered by M, then the conditions above translate to the constraints:

If $\omega \in \Omega^+_{BS}$ then

$$\sum_{\substack{j=1 \\ \omega \in M^+_j}}^{p} r_j \geq 1, \text{ and}$$

if $\omega \in \Omega^-_{BS}$ then

$$\sum_{\substack{h=1 \\ \omega \in M^-_h}}^{q} s_h \geq 1.$$

Obviously, the replacement of 1 in the right hand side of the above inequalities by a positive integer d, as in (2), can lead to an increase of the robustness of resulting model.

There are several ways of defining an appropriate objective function. The simplest definition would require only the minimization of the number of patterns used, i.e.

$$\text{minimize } \sum_{j=1}^{p} r_j + \sum_{h=1}^{q} s_h. \tag{3}$$

An alternative point of view may require the identification of a family of patterns in which the overlap between positive and negative patterns is minimized. It is clear from the definition of patterns that there are no points in Ω_{BS}, which can belong at the same time to a positive and a negative pattern. It may happen, however, that a new point may be covered simultaneously by a positive and a negative pattern.

For each pair consisting of a positive and a negative pattern it is easy to calculate the number of points covered by their intersection. Indeed, if the definition of a positive and a negative pattern "conflicts", i.e. there is a variable x appearing uncomplemented in the definition of one of the two patterns, and appearing complemented in the definition of the

other one, then the subcubes defined by the two patterns are disjoint. If, however, the definitions of the two patterns do not conflict, and if the number of variables appearing in the two patterns is respectively α and β, with γ variables common to both patterns, then clearly the number of points in the intersection of the subcubes generated by these two patterns is $2^{t-\alpha-\beta+\gamma}$. If we define

$$\delta_{j,h} = \begin{cases} 2^{t-\alpha-\beta+\gamma} & \text{if patterns } M_j^+ \text{ and } M_h^- \text{ do not conflict;} \\ 0 & \text{otherwise,} \end{cases}$$

then the quadratic objective defining this problem is to

$$\text{minimize} \sum_{j=1}^{p} \sum_{h=1}^{q} \delta_{j,h}\, r_j\, s_h. \tag{4}$$

Table 6 presents data concerning the comparative advantages and disadvantages of using the complete model (consisting of all the maximum patterns), a model using the minimum number of maximum patterns (obtained by solving exactly the set covering problem whose objective function is (3)), and the overlap minimizing quadratic set covering problem (solved by using the L_2-best linear approximation of its quadratic objective function (4)). Comparing the number of patterns used in the three models, it can be seen that the second and third models use about half of the number of patterns used by the first one, and do not differ significantly. It is very interesting to note that on the average the second model performs almost as well as the first one (the average difference being of about 2%), while the third model's accuracy is only 1% below that of the second one. It seems clear from these results that the use of the first model is much less advantageous than that of the second and third. When deciding about the use of the second or third model, perhaps the major factor is the number of unclassified observations, which turned out to be slightly smaller in the quadratic model.

	Number of Patterns			Accuracy		
Dataset	Complete Model	Min. Size Model	Approx. Min. Overlap Model[4]	Complete Model	Min. Size Model	Approx. Min. Overlap Model[4]
BCW	55.0	25.2	25.5	95.0%	92.0%	92.6%
HD	82.0	42.5	43.7	83.4%	80.9%	80.2%
PID	129.1	53.2	55.4	77.2%	77.3%	73.7%
BLD	140.9	60.0	61.5	74.9%	72.5%	72.2%
Average	101.8	45.2	46.5	82.6%	80.7%	79.7%

Table 6: Comparison between the three approaches for model formation.

[4]Using L_2-best linear approximation of the quadratic model.

2.5 Accuracy and Validation

A typical component of the analysis of medical data is the validation or cross-validation of conclusions. When the original dataset is sufficiently large to allow the partition of the observations into a "training" and a "test set", the first one is used to derive a mathematical model and draw conclusions from it, while the second one is used to test the validity of the conclusions derived in this way. Because of the difficulty in working with large sets of patients displaying certain conditions, the medical datasets consist frequently of relatively small sets of observations. In view of this fact, cross-validation techniques are frequently used for evaluating the quality of conclusions derived from the analysis of medical data.

The most frequently used cross-validation technique is the usual *k-folding* method of statistics. This method consists in the random partitioning of the set of observations into k (approximately) equally sized subsets; one of these subsets is designated as the "test set", a model is built on the union of the remaining $k-1$ subsets (which form the "training set") and then tested on the k-th subset. This process is repeated k times by changing the subset taken as test set, and the average accuracy is then reported as a quality measure of the proposed method. In the results presented in this paper we have usually taken k to be 10, and reported as accuracy the average of 10 to 20 k-folding experiments.

A special case of k-folding is the so-called "jackknifing", or "leave-one-out" technique, in which k is taken to be equal to the number of observations in the dataset, i.e. the test sets consist always of a single patient.

The classification of a new patient by LAD can either lead to the prediction of he or she being positive, or negative, or "unclassified" (although the last category is usually very small). We shall define the *accuracy* of predictions to be simply the proportion of correctly classified patients in the test set.

Two other concepts which are frequently used in medicine are those of *sensitivity* and *specificity*. *Sensitivity* is simply the proportion of correctly classified positive observations within the set of positive observations in the test set. Similarly, *specificity* is simply the proportion of correctly classified negative observations within the set of negative observations in the test set.

Finally, the *hazard ratio* of a set Σ of observations, another frequently used quality measure in medicine is the proportion of two proportions: the proportion of positive observations in the set Σ, and the proportion of positive observations in the complement of Σ. Usually, Σ is taken to be the set of observations in the test set which are predicted to be positive by LAD or other data analysis method.

3 Applications to Medicine

We shall present in this section several results obtained by applying LAD to different typical problems arising in medicine.

3.1 Differential IIP Diagnosis Using Radiological Data[5]

Idiopathic interstitial pneumonias (*IIPs*) are a group of disorders resulting from damage to the lung parenchyma by varying patterns of inflammation and fibrosis. Various forms of IIP differ both in their prognoses and their therapies, but are not easily distinguishable using clinical, biological and radiological data, and therefore frequently requiring pulmonary biopsies to establish the diagnosis.

In order to avoid the difficulties related to biopsy, we have applied LAD to computed tomography (CT) data, to distinguish between 3 types of IIPs: *Idiopathic Pulmonary Fibrosis (IPF)*, *Non Specific Interstitial Pneumonia (NSIP)*, and *Desquamative Interstitial Pneumonia (DIP)*.

One of the characteristics of this dataset was its extremely small size: it consisted only of 56 patients, and included 13 attributes. In spite of the very small size of the dataset, some surprisingly strong patterns, with prevalences ranging between 40% and 86% have been identified in it.

In a first step, LAD identified 2 outliers, this finding having been confirmed later by the medical records of those 2 patients; one of them turned actually out to have been exposed to asbestos, while the other one's lung pathology data was found to be atypical in all features.

In contrast to the difficulties encountered by experienced medical researchers to distinguish between the 3 types of interstitial pneumonia, the LAD study allowed the precise diagnosis of 44 of the 54 remaining patients, made errors in 6 cases, and left unclassified the other 4 patients.

The diagnosis accuracy of over 80% obtained by LAD is far superior to those reported recently [24, 25] in the medical literature (correct diagnoses in 32% to 70% of IPF cases, 60% of DIP, and only 9% of NSIP).

Beside diagnosis the study identified several variables as having a blocking or a promoting effect on some forms of interstitial pneumonia.

The encouraging results of this investigation form the basis of a forthcoming study of a broader population of IIPs, which will include not only CT data, but also clinical and biological ones.

3.2 Ovarian Cancer Diagnosis Using a Large Proteomic Dataset[6]

Petricoin et al. published in 2002 [27] the results of a successful experiment in the diagnosis of ovarian cancer based on the analysis of mass spectroscopy generated data containing expression profiles of 15,154 peptides defined by their mass per charge (m/z) ratios in serum of 162 ovarian cancer and 91 control cases. The high level of interest of this investigation was demonstrated by the New York Times' prompt release [1] of an announcement of the essential findings of this study.

[5]Based on a collaborative study with researchers from Hôpital Avicenne (Paris), University of Grenoble, and RUTCOR, reported in [15].

[6]Based on a collaborative study with researchers from the National Institutes of Health, the Food and Drug Administration (FDA), the Cancer Institute of New Jersey, and RUTCOR, reported in [8].

Using LAD we have re-examined the Petricoin-Liotta dataset[7], and its subsequently revised versions, and identified in them 3 subsets consisting respectively of 7, 8 and 9 peptides chosen from the 15,154 peptides in the dataset. The 9 peptides found by LAD have relatively low correlation coefficients with the outcome, in marked contrast with the widely accepted idea of basing the selection of biomarkers on their individual distinguishing power.

An interesting finding of the study is the existence of very simple "combinatorial biomarkers". For example, in 97% of the positive cases the expression of the intensity level of the peptide with the m/z value 235.8296 is *low* (i.e. below a certain prescribed threshold) while the expression of the intensity level of the peptide with m/z value 435.46452 is *high* (i.e. above a certain prescribed threshold). Moreover, this combination of intensity levels does not occur in any of the negative cases. Thus, each of the patterns can be viewed as a logically synthesized combinatorial biomarker.

Three different diagnostic models consisting of such powerful patterns (combinatorial biomarkers) have been built on these support sets, and shown by systematic cross-validation experiments to have sensitivities ranging between 96.7% and 100%, and specificities ranging from 95.1% to 100%. Both the sensitivity and the specificity of the proposed "complete model", which involves all the 9 peptides (selected from the 15,154), are of 100% each.

The high accuracy of these diagnostic models indicates clearly the presence of distinctive differences in the proteomic serum spectra from patients with ovarian cancer compared to unaffected patients, and it fully reconfirms the essence of the conclusions of [27].

One of the most important problems in ovarian cancer is its detection in stage I, when the possibility of treating it by surgery alone, without the need of chemotherapy, provides superior survival rates. Using again the LAD approach, we have detected a support set of 6 peptides on which we have built several diagnostic models for recognizing stage I ovarian cancer. Both the average sensitivity and specificity of the "complete model" built on this support set turned out in 120 cross-validation experiments to be of 100%.

Another interesting conclusion of this study is that using only peptides with relatively high m/z values does not allow the construction of accurate diagnostic models. It is not clear at this stage whether this phenomenon is due to imprecisions in measurements, or to biological reasons.

An additional result of the study is the identification of several large and coherent groups of cases with strikingly similar combinatorial characteristics, raising significant questions about their possible biological similitudes.

3.3 Genomic Data-based Breast Cancer Prognosis[8]

In a recent study [29] van't Veer et al. proposed to predict the clinical outcome of breast cancer (i.e. to identify the cases which will develop metastases within 5 years) based on the analysis of gene expression signatures. The importance of this problem is due to the fact

[7] See http://clinicalproteomics.steem.com.

[8] Based on a collaborative study with researchers from the Robert Wood Johnson Medical School, Rutgers University's Department of Biology, and RUTCOR, reported in [4].

that the available adjuvant (chemo or hormone) therapy, which reduces by about one third the risk of distant metastases is not really necessary for three quarters of the patients who currently receive it, and can even have serious side effects.

The attributes in the van't Veer study[9] correspond to more than 25,000 human genes, while the number of patients is only 97. The attributes are measured by the fluorescence intensities of RNA hybridized to microarrays of oligonucleotides. The 97 lymph-node-negative breast cancer patients are grouped into a training set of 78 and a test set of 19 cases. The training set includes 34 positive cases (having a "poor prognosis" signature, i.e. having less than 5 years of metastasis-free survival), and 44 negative cases (having a "good prognosis" signature, i.e. having more than 5 years metastasis-free survival). The test set includes 12 positive and 7 negative cases. The van't Veer study identified 231 biomarkers of metastasis (all having large correlations with the outcome), and an optimal prognosis classifier based on 70 genes, and having an accuracy of 83.3% on the training set, and 89.5% on the test set.

By applying LAD to the van't Veer data, we have identified a support set of 16 genes, which includes only 2 of the genes appearing in the van't Veer study. On this support set of 16 genes we have identified 39 positive and 93 negative patterns. On the training set the accuracy of the proposed prognostic system consisting of these 142 patterns is of 100%, while on the test set it classifies correctly 18 of the 19 cases (accuracy of 94.7%).

One of the genes appearing in almost 40% of the significant patterns was shown to be a highly significant biomarker. Along with this gene, 6 promoters and 10 blockers have been identified in the support set. It is interesting that each of the genes in the support set is either a contributor or a blocker, which is a very unusual situation, since most datasets contain very few attributes with a consistent (monotone) behavior.

A new subclass of positive patients, containing 13 cases has also been discovered. The patients in this group have a fully predictable behavior (the sensitivity in this group is 100%, compared to 81% over the entire set of positive cases), have distinctive gene expression ranges, and several other special characteristics.

An interesting conclusion of the study is that the patients included in the training set and the test set turned out to have differing characteristics. Surprisingly, the accuracy in the test set (94.7%) turned out to be even higher than that estimated by cross-validation on the training set (85.9%). Several patterns have been identified, each of which including only one or two genes, and providing a 100% distinction between the training and the test sets. Moreover, the genes in the support set define an interval in the 16-dimensional real space, which includes all the 19 test cases and none of the training cases, providing thus a complete separation of the training and the test sets.

3.4 Logical Analysis of Diffuse Large B-Cell Lymphomas[10]

Diffuse large B-cell lymphoma (DLBCL) is one of the most common subtypes of non-Hodgkin lymphoma (NHL), accounting for 31% of NHL cases. Using modern chemotherapy, 50%

[9] See http://www.rii.com/publications/2002/vantveer.html.

[10] Based on a collaborative study with researchers from the Robert Wood Johnson Medical School, Rutgers University's Department of Biology, and RUTCOR, reported in [5].

of patients achieve a long-term, disease-free survival. Recently Shipp and coworkers [28] described the use of a correlation-based, supervised learning technique (i) to distinguish DLBCLs from follicular lymphomas (FL), and (ii) to predict the clinical course of cases of DLBCL (i.e. to distinguish between poor and good prognosis cases).

The goal of this study was to reexamine the same two problems with the help of LAD, using the microarray dataset of Shipp et al.[11], which contains the intensity levels of 6,817 genes of 58 patients with diffuse large B-cell lymphoma (DLBCL) and 19 with follicular lymphoma (FL). Out of the 58 patients with DLBCL in the lymphoma dataset [w4] 26 had poor prognoses, and 32 had good prognoses.

Results for problem (i): For the differential diagnosis of DLBCL vs. FL, a model based on 8 significant genes was constructed and shown to have a sensitivity of 94.7% and a specificity of 100% on the test set. It is interesting to remark that in spite of the fact that the correlation of the expression levels of none of the 6,817 genes in the dataset with the DLBCL vs. FL outcome exceeded 62% in absolute value, 37 extremely powerful patterns have been found, each involving the expression levels of only 2 of the genes in a support set of 20 genes; what makes this collection remarkable is that each of these patterns had prevalence in excess of 90%, either in the DLBCL or the FL class, and 0% in the other class.

Among the biological conclusions derived from the analysis of the most powerful patterns, it is worth mentioning that more than 75% of the patterns with large prevalences included a gene belonging to the "cell surface proteins and receptors" class, while half of the significant patterns included a gene belonging to the "DNA replication, combination and repair" class. Most importantly, every single pattern included at least one gene belonging to one of these two classes.

The existence of such powerful patterns has made the construction of extremely accurate classification models possible. A model built on a support set of only 8 genes, and using only 13 of the patterns identified in this dataset, had a sensitivity of 94.7% and a specificity of 100% on the test set, while on the training set both the sensitivity and the specificity were of 100%.

Results for problem (ii): Similarly to problem (i), 75 very powerful patterns, having prevalences in excess of 72% and 86% respectively have been identified in the poor, respectively good, prognosis classes. A classification model consisting of 16 patterns was built on a support set of 8 genes; the sensitivity and specificity of this model on test sets were of 87.5% and of 90% respectively, while on the training set both the sensitivity and the specificity were of 100%.

One of the most important conclusions of this study concerns the interactions among genes, illustrated by the presence of powerful combinatorial biomarkers (patterns), and requiring biological explanation. Beside the identification of several genes of prominent importance in the two problems, e.g. butyrophilin(BTF1)mRNA, whose expression level appears in more than two thirds of the patterns in problem (ii), several other genes with clear strongly monotonic (i.e. promoting or blocking) properties have been discovered.

Consideration of the function and location of the butyrophilin product, and of several

[11]See http://www.genome.wi.mit.du/MPR/lymphoma.

other support set gene products that appear with it in the various patterns, suggests a possible hypothesis for their relationship in a transduction pathway, in which an extracellular ligand interacts with a membrane receptor, transduces a signal via a cascade to the nucleus, and influences DNA replication. The biological confirmation of this, and of many other hypotheses which can be derived from this study, represent examples of challenging open biomedical research problems, inspired by the mathematical/computatioal analysis of data.

3.5 Coronary Risk Prediction[12]

The objective of this collaborative risk stratification study carried out with a team of medical researchers at the Cleveland Clinic Foundation was to distinguish within a population of patients with known or suspected coronary artery disease, groups of patients at high or low mortality rates. The study was based on Cleveland Clinic Foundation's data, which includes 9,454 patients, of whom 312 died during an observation period of 9 years. 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), as well as an indication of whether the patient died during the observation period.

In most applications of LAD and of other data analysis techniques the representation of the analyzed datasets in real space admitted a more or less "crispy" separation into homogeneous zones, containing only positive or only negative points. The disproportionate sizes of the two groups of patients in the study, and the "inseparable" character of the dataset, have prompted a new definition of the positive and negative classes (see Figure 3).

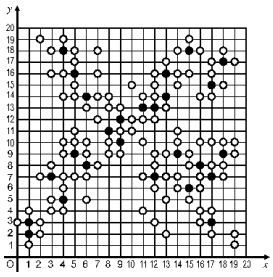


Figure 3: An inseparable dataset.

A group of patients was defined as "positive" or "high risk" if its mortality rate exceeded at least five times the average mortality rate of the entire population (3.3%), i.e. if it was at least 16.5%. Similarly, a group of patients was defined as "negative" or "low risk" if its mortality rate was less than one fifth of the average mortality rate of the entire population, i.e. if it was less than 0.66%. A patient was defined to be at high risk (respectively, low

[12]Based on a collaborative study with researchers from the Cleveland Clinic Foundation, and RUTCOR, reported in [9] and [26].

		LAD Index	
		High	Low
Cox Score	High	$3.99 \times \mu$	$0.92 \times \mu$
	Low	$1.47 \times \mu$	$0.39 \times \mu$

Table 7: LAD Index vs. Cox Score (Average mortality: $\mu = 3.3\%$).

risk) if its measurements satisfied some positive (respectively, negative) patterns, but none of the negative (respectively, positive) ones. Patients satisfying both positive and negative patterns were classified on the basis of the corresponding signs of a discriminant. The value associated by this discriminant to a patient was called the patient's *Prognostic Index*.

In the cardiovascular literature, risk stratification schemes are typically based on standard statistical models, such as logistic regression or Cox proportional hazards. A common problem with these approaches is that, although high risk patients can be easily identified, they account usually only for a minority of subsequent clinical events. 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. The ideal risk stratification scheme has to identify a small subset of patients who will in fact account for the vast majority of deaths.

Using the Prognostic Index, the number of patients classified into the high risk and low risk categories were respectively of 20% and 77%. Moreover, 75% of the patients who died during the observation period belonged to the high risk category defined by LAD.

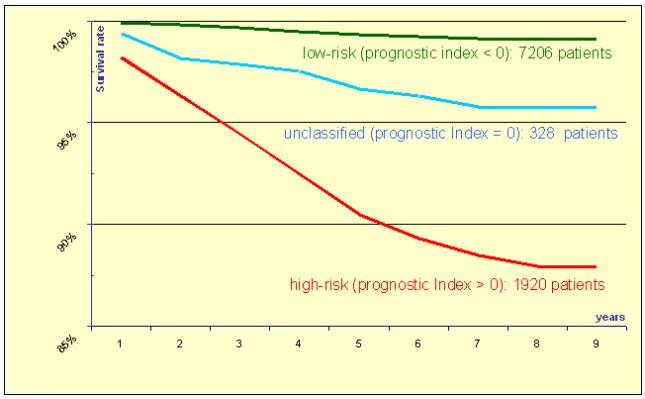


Figure 4: Survival dynamics of high and low risk patients.

In comparing the LAD-based Prognostic Index to the Cox Score, widely used by cardiologists, it was shown first that the classifications given by both indicators agree in three out of four cases, and that the Prognostic Index outperforms the Cox Score slightly but consistently (see Table 7). It should also be noted that the correlation of the Prognostic Index with the mortality rate of patients is of 84%.

The dynamics of survival of patients classified by LAD as being at high risk (respectively, low risk) is illustrated in Figure 4.

3.6 Cell Growth Prediction for Polymeric Biomaterial Design[13]

Since extremely large combinatorial biomaterial libraries are being developed in many laboratories, and since the experimental evaluation of the qualities of various polymers in these libraries is both expensive and time-consuming, the need for the development of analytic procedures for the pre-selection of those polymers that should actually be tested is becoming more and more important. In particular, computational models that can predict the cellular response to implanted biomaterials (e.g. artificial bones) can play an invaluable role in the design of medical devices whose functions depend on controlling cell-material interaction at the device surface.

As a step in this direction we have examined the properties of a library of 112 polymers on growth of two different cell types: rat lung fibroblasts (RLF) and normal foreskin fibroblasts (NFF, human cells). LAD was trained on a subset of 62 polymers and was then used to predict cell growth on some of 50 untested polymers taken from that library.

LAD found patterns of structural and physical parameters, which led to the classification of each polymer as high, medium or low cell growth substratum, and provided specific correlations between cell growth on the one hand, and chemical composition, bulk properties, surface chemistry on the other.

One of the most interesting outcomes of this study was that a group of 6 "superior" polymers within the training set was found to satisfy not just one or two, but all patterns of polymer properties associated to high NFF cell growth. From the 6 superior polymers a single new pattern was established which describes completely and exclusively these 6 materials; this special pattern is expressed in terms of restrictions imposed on the pendant chain used, backbone, glass transition, contact angle, and flexibility index. In this way the LAD model gave rise to surprising explicit design criteria for the development of polymers that will support the growth of RLF cells.

A similar approach was used to identify the "inferior" polymers, i.e. those with low cell growth within the training set. One specific pattern was constructed which is capable of identifying all low cell growth polymers for RLF cells. Again, LAD resulted in remarkably precise design criteria for polymeric substrata that could be used in applications where cell growth is undesirable.

The growth behavior of NFF cells was more complex than that of the RLF cell line, and therefore the LAD-derived criteria for polymeric substrata for NFF cells where less definite. We expect that the refinement of the present study by the inclusion of additional input parameters will allow us to address this point in more detail.

The models were tested experimentally by using them to predict cell growth on the remaining 50 previously untested polymers in the library, some of which were actually synthesized for experimental validation. The cell culture results showed that the LAD model found optimal ranges for polymer chemical composition, surface chemistry, and bulk properties. Moreover, it classified correctly the approximate range of cell growth for 83% of the

[13]Based on a collaborative study with researchers from the New Jersey Center for Biomaterials, Rutgers University's Department of Chemistry and Chemical Biology, and RUTCOR, reported in [2].

polymers tested in the case of RLF cells, and 73% for the NFF cells. Particular noteworthy is that LAD correctly identified high performing polymer surfaces, identifying promising "lead" polymers for applications that require high or low cell growth.

To our knowledge this research represents the first time a computer model based on the recognition of patterns of polymer composition and properties has been used to predict cell growth outcomes on previously untested biomaterials. It is expected that the patterns identified by LAD will be of substantial assistance in (i) eliminating the need of synthesizing polymers of low expected value, and (ii) directing the experimental stage of design towards the synthesis of the most promising biomedical polymers.

References

[1] Ciphergen Confirms Use of Its ProteinChip(R) System in Ovarian Cancer Study Published in the Lancet. *The New York Times*, February 11, 2002.

[2] Abramson S., G. Alexe, P.L. Hammer, D. Knight and J. Kohn. A Computational Approach to Predicting Cell Growth on Polymeric Biomaterials. *Journal of Biomedical Material Research*, (forthcoming).

[3] Alexe G., P.L. Hammer. Spanned Patterns in the Logical Analysis of Data. *Discrete Applied Mathematics (to appear).*

[4] Alexe G., S. Alexe, D. Axelrod, E. Boros, P.L. Hammer, M. Reiss. Combinatorial Analysis of Breast Cancer Data from Gene Expression Microarrays. *RUTCOR Technical Report*, 2004.

[5] Alexe G., S. Alexe, D. Axelrod, P.L. Hammer, D. Weissmann. Logical Analysis of Diffuse Large B-Cell Lymphomas. *Artificial Intelligence in Medicine*, 2005 (forthcoming).

[6] Alexe G., S. Alexe, P. L. Hammer. Pattern-Based Clustering and Attribute Analysis. *RUTCOR Research Report 10-2003 (to appear in Soft Computing)*, 2003.

[7] Alexe G., S. Alexe, P.L. Hammer, B. Vizvari. Pattern-Based Feature Selections in Genomics and Proteomics. *RUTCOR Research Report 7-2003*, 2003.

[8] Alexe G., S. Alexe, P.L. Hammer, L. Liotta, E. Petricoin, and M. Reiss. Ovarian Cancer Detection by Logical Analysis of Proteomic Data. *Proteomics*, 4(3):766–783, 2004.

[9] Alexe S., E. Blackstone, P.L. Hammer, H. Ishwaran, M. Lauer, C. Pothier Snader. Coronary Risk Prediction by Logical Analysis of Data. *Annals of Operations Research*, 119:15–42, 2003.

[10] Alexe S., P.L. Hammer. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. *Discrete Applied Mathematics (to appear).*

[11] Axelrod D., T. Bonates, P. L. Hammer, I. Lozina. From Diagnosis to Therapy via LAD. *Invited Lecture at INFORMS Annual Meeting, Denver, CO*, October, 2004.

[12] Bonates T., P. L. Hammer, A. Kogan. Maximum Patterns in Datasets. *RUTCOR Research Report*, 2005.

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

[14] Boros E., 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(2):292–306, 2000.

[15] Brauner M.W., N. Brauner, P.L. Hammer, I. Lozina, D. Valeyre. Logical Analysis of Computed Tomography Data to Differentiate Entities of Idiopathic Interstitial Pneumonias. *Data Mining in Biomedicine, Biocomputing, Springer*, (forthcoming).

[16] Crama Y., P.L. Hammer, T. Ibaraki. Cause-effect Relationships and Partially Defined Boolean Functions. *Annals of Operations Research*, 16:299–325, 1988.

[17] Eckstein J., P.L. Hammer, Y. Liu, M. Nediak, B. Simeone. The Maximum Box Problem and its Application to Data Analysis. *RUTCOR Research Report 4-2002. To appear in Computational Optimization and Applications*.

[18] Hammer A., P.L. Hammer, I. Muchnik. Logical Analysis of Chinese Labor Productivity Patterns. *Annals of Operations Research*, 87:165–176, 1999.

[19] Hammer P.L. The Logic of Cause-effect Relationships. Lecture at the International Conference on Multi-Attribute Decision Making via Operations Research-bsaed Expert Systems, Passau, Germany, 1986.

[20] Hammer P.L., A. Kogan, B. Simeone, S. Szedmak. Pareto-Optimal Patterns in Logical Analysis of Data. *RUTCOR Research Report, 7-2001, and Discrete Applied Mathematics (in print)*, 2001.

[21] Hammer P.L., A. Kogan, M. Lejeune. Country Risk Ratings: Statistical and Combinatorial Non-recursive Models. *RUTCOR Research Report 8-2004*, 2004.

[22] Hammer P.L., A. Kogan, M. Lejeune. Modeling Country Risk Ratings Using Partial Orders. *RUTCOR Research Report 24-2004*, 2004.

[23] Hammer P.L., R. Holzman. Approximations of Pseudo-Boolean Functions; Applications to Game Theory. *Methods and Models of Operations Research*, 39:3–21, 1992.

[24] Hartman T.E., S.J. Swensen, D.M. Hansell, T.V. Colby, J.L. Myers, H.D. Tazelaar, A.G. Nicholson, A.U. Wells, J.H. Ryu, D.E. Midthun, R.M. du Bois, N.L. Muller. Nonspecific Interstitial Pneumonia: Variable Appearance at High-Resolution Chest CT. *Radiology*, 217(3):701–705, 2000.

[25] Johkoh T., N.L. Muller, Y. Cartier, P.V. Kavanagh, T.E. Hartman, M. Akira, K. Ichikado, M. Ando, H. Nakamura. Idiopathic Interstitial Pneumonias: Diagnostic Accuracy of Thin-Section CT in 129 Patients. *Radiology*, 211(2):555–560, 1999.

[26] Lauer M., S. Alexe, E.H. Blackstone, P.L. Hammer, H. Ishwaran, C. Pothier Snader. Use of the Logical Analysis of Data Method for Assessing Long-Term Mortality Risk After Exercise Electrocardiography. *Circulation*, 106:685–690, 2002.

[27] Petricoin E.F., A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta. Use of Proteomic Patterns in Serum to Identify Ovarian Cancer. *The Lancet*, 359:572–577, 2002.

[28] Shipp M.A., K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning. *Nature Med.*, 8(1):68–74, 2002.

[29] van 't Veer L.J., H. Dai, M.J. van De Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, S.H. Friend. Gene Expression Profiling PredictsClinical Outcome of Breast Cancer. *Nature*, 12:292–306, 2002.

[30] Witten I.H., E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 2000.