R U T C O R
R ESEARCH
R E P O R T

LOGICAL ANALYSIS OF COMPUTED TOMOGRAPHY DATA TO DIFFERENTIATE ENTITIES OF IDIOPATHIC INTERSTITIAL PNEUMONIAS

M.W. Brauner[a] N. Brauner[b] P.L. Hammer[c]
I. Lozina[c] D. Valeyre[d]

RRR 30-2004, SEPTEMBER 2004

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] Department of Radiology, Fédération MARTHA, UFR Bobigny, Université Paris 13 et Hôpital Avicenne AP-HP, 125, route de Stalingrad, 93009 Bobigny Cedex, France.
E-mail: michel.brauner@wanadoo.fr
[b] Laboratoire Leibniz-IMAG, 46 av. Felix Viallet, 38031 GRENOBLE Cedex, France.
E-mail: Nadia.Brauner@imag.fr
[c] RUTCOR, Rutgers University, 640 Bartholomew Rd., Piscataway NJ, 08854-8003 USA.
E-mail: hammer@rutcor.rutgers.edu, ilozina@rutcor.rutgers.edu
[d] Department of Pneumology, Fédération MARTHA, UFR Bobigny, Université Paris 13 et Hôpital Avicenne AP-HP, 125, route de Stalingrad, 93009 Bobigny Cedex, France.

RUTCOR RESEARCH REPORT
RRR 30-2004, SEPTEMBER 2004

LOGICAL ANALYSIS OF COMPUTED TOMOGRAPHY DATA TO DIFFERENTIATE ENTITIES OF IDIOPATHIC INTERSTITIAL PNEUMONIAS

M.W. Brauner N. Brauner P.L. Hammer I. Lozina D. Valeyre

Abstract. The aim of this paper is to analyze computed tomography (CT) data by using the Logical Analysis of Data (LAD) methodology in order to distinguish between three types of idiopathic interstitial pneumonias (IIPs). The paper demonstrates that LAD can distinguish with high accuracy different forms (IPF, NSIP and DIP) of IIPs. It shows also that the patterns developed by LAD techniques provide additional information about outliers, redundant features, the relative significance of attributes, and makes possible the identification of promoters and blockers of various forms of IIPs.

Acknowledgements: Peter L. Hammer and Irina Lozina gratefully acknowledge the partial support of NSF through Grant # NSF-IIS-0312953.

1 Introduction

The idiopathic interstitial pneumonias (IIPs) are a heterogeneous group of nonneoplastic disorders resulting from damage to the lung parenchyma by varying patterns of inflammation and fibrosis. A new classification of IIPs was established in 2001 by an International Consensus Statement defining the clinical manifestations, pathology and radiological features of patients with IIPs ([4]). 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. The aim of this paper is to analyze computed tomography (CT) data by techniques of biomedical informatics to distinguish between 3 types of IIPs:

- Idiopathic Pulmonary Fibrosis (IPF)
- Non Specific Interstitial Pneumonia (NSIP)
- Desquamative Interstitial Pneumonia (DIP)

2 Patients and methods

This is a study of the CT scans in patients with IIPs referred to the Department of Respiratory Medicine, Avicenne Hospital, Bobigny, France, for medical advice on diagnosis and therapy. The diagnosis was established on clinical, radiographic and pathologic (i.e. biopsy–based) data. The 56 patients included 34 IPFs, 15 NSIPs, and 7 DIPs.

We reviewed the CT examination of the chest from these patients. CT scans were evaluated for the presence of signs and a score was established for the 2 main lesions, ground-glass attenuation and reticulation. Pulmonary disease severity on thin section CT scans was scored semiquantitatively in upper, middle and lower lung zones. The six areas of the lung were defined as follows: the upper zones are above the level of the carina; the middle zones, between the level of the carina and the level of the inferior pulmonary veins; and the lower zones, under the level of the inferior pulmonary veins. The profusion of opacities was recorded separately in the six areas of the lung to yield a total score of parenchymal opacities. The severity was scored in each area according to four basic categories : 0 = normal, 1 = slight, 2 = moderate, 3 = advanced (total: 0-18)

The data consisted of the binary attributes 1, 2, …,10, and the numerical attributes 11, 12, and 13 listed bellow:

1. IIT intralobular interstitial thickening
2. HC honeycombing
3. TB traction bronchiectasis
4. GG1 ground-glass attenuation
5. BRVX peri-bronchovascular thickening
6. PL polygonal lines
7. HPL hilo-peripheral lines

8. SL	septal lines
9. AC	airspace consolidation
10. N	nodules
11. GG2	ground-glass attenuation score
12. RET [1]	reticulation score
13. GG2/RET	ground-glass attenuation/reticulation score

The analysis of this dataset was carried out using the combinatorics, optimization and logic based methodology called the *Logical Analysis of Data* (LAD) proposed in ([7], [8]). Detailed description of this methodology appears in ([5]). Also, a brief outline of LAD appears in this volume (in [3]). Among previous studies dealing with applications of LAD to medical problems we mention ([1], [2], [9]).

The choice of LAD for analyzing the IIP data is due on the one hand to its proven possibility to provide highly accurate classifications, and on the other hand to the usefulness of LAD patterns in analyzing the significance and nature of attributes.

The conclusions of LAD have been confirmed by other methods used in bioinformatics (neural networks, decision trees, support vector machine, etc.). An additional result of the study was the identification by LAD of two outliers, which turned out to have complete medical explanation.

3 Outliers

We have constructed 3 different LAD models to distinguish between IPF, NSIP or DIP patients:

- model I to distinguish IPF patients (considered to be the positive observations in this model) from non- IPF patients (negative observations);
- model II to distinguish NSIP patients (positive in this model) from non- NSIP patients (negative observations);
- model III to distinguish DIP patients (positive in this model) from non-DIP patients (negative observations).

These models use only pure patterns, their degrees are at most 4, and their prevalences range between 40% and 85.7%.

3.1 Two suspicious observations.

The classification given by the 3 LAD models for the 56 observations in the dataset is shown in Table 1. It can be seen that all the 56 classifications are correct, but only 54 of them are precise. In fact the classifications of the observations s003 and s046 are vague. Since observation s003 turns out to be classified as being either a DIP or an NSIP patient, we have built an additional model to distinguish between these two classes. It turns out that the model contains only one pattern covering observation s003. This pattern shows (correctly) that s003 is a DIP

[1] RET is a generic term which includes the three main fibrotic lesions : ITT, HC and TB

Table 1

Observations	Given Classification	Classification by LAD Models			
		IPF/non-IPF	NSIP/nonNSIP	DIP/nonDIP	Conclusion
s001	DIP	0	?	1	DIP
s002	DIP	0	0	1	DIP
s003	DIP	0	?	?	NSIP or DIP
s004	DIP	0	0	1	DIP
s005	DIP	0	?	1	DIP
s006	DIP	0	?	1	DIP
s007	DIP	0	0	1	DIP
s008	IPF	1	0	0	IPF
s009	IPF	1	?	0	IPF
s010	IPF	1	?	0	IPF
s011	IPF	1	0	0	IPF
s012	IPF	1	0	0	IPF
s013	IPF	1	0	0	IPF
s014	IPF	1	0	0	IPF
s015	IPF	1	0	0	IPF
s016	IPF	1	?	0	IPF
s017	IPF	1	?	0	IPF
s018	IPF	1	0	0	IPF
s019	IPF	1	0	0	IPF
s020	IPF	1	0	0	IPF
s021	IPF	1	?	0	IPF
s022	IPF	1	0	0	IPF
s023	IPF	1	0	0	IPF
s024	IPF	1	0	0	IPF
s025	IPF	1	0	0	IPF
s026	IPF	1	0	0	IPF
s027	IPF	1	0	0	IPF
s028	IPF	1	0	0	IPF
s029	IPF	1	0	0	IPF
s030	IPF	1	0	0	IPF
s031	IPF	1	0	0	IPF
s032	IPF	1	0	0	IPF
s033	IPF	1	0	0	IPF
s034	IPF	1	0	0	IPF
s035	IPF	1	0	0	IPF
s036	IPF	1	0	0	IPF
s037	IPF	1	0	0	IPF
s038	IPF	1	0	0	IPF
s039	IPF	1	0	0	IPF
s040	IPF	1	0	0	IPF
s041	IPF	1	0	0	IPF
s042	NSIP	0	1	0	NSIP
s043	NSIP	0	1	0	NSIP
s044	NSIP	0	1	0	NSIP
s045	NSIP	0	1	0	NSIP
s046	NSIP	?	?	0	IPF or NSIP
s047	NSIP	0	1	0	NSIP
s048	NSIP	0	1	?	NSIP
s049	NSIP	0	1	?	NSIP
s050	NSIP	0	1	0	NSIP
s051	NSIP	0	1	0	NSIP
s052	NSIP	0	1	0	NSIP
s053	NSIP	0	1	0	NSIP
s054	NSIP	0	1	0	NSIP
s055	NSIP	0	1	0	NSIP
s056	NSIP	0	1	0	NSIP

patient, however it does not cover any other observation, i.e. its prevalence is so low that it cannot be considered reliable. A very similar argument concerning the observation s046 shows that in a model distinguishing IPF/NSIP cases, it is classified as being an NSIP case, however this classification is based only on extremely weak patterns, whose reliability is low. The facts signaled above, raise suspicions about the specific nature of these two observations, and raise the question of whether they should be included at all in the dataset.

3.2 Medical confirmation

In view of the suspicions related to these two observations, the medical records of these two patients have been re-examined. It was found that patient s003 was exposed to asbestos, and therefore its classification as DIP is uncertain. Asbestosis may be responsible for a pathologic aspect similar to that of IPF, but very different from DIP. It is also possible that the pathologic result on the biopsy of a very small area of the lung was wrong. Also, it was found that the data of patient s046 are highly atypical in all the features (age, clinical data and lung pathology), and it was suggested that in view of these reasons, the patient should be considered unclassable and removed from the dataset.

3.3 Improving classification accuracy by removing outliers

The medical confirmation of the suspicions raised by the inability of the LAD models to classify the two unusual observations, have led us to check the ways in which the accuracy of various classification methods changes when these two observations are removed from the dataset. In order to evaluate these changes, we have applied five classification methods taken from the WEKA package ([*http: //www.cs.waikato.ac.nz/~ml/weka/index.html*]), separately to the original dataset of 56 observations, and to the dataset of 54 observations obtained by removing the two suspicious ones. The 5 methods used for this purpose were: artificial neural networks ("Multilayer Perceptron" in WEKA), linear logistic regression classifier ("Simple Logistic" in WEKA), support vector machine classifier ("SMO" in WEKA), nearest-neighbor classifier ("IB1" in WEKA), and decision trees ("J48" in WEKA).

Twenty 3-folding experiments were carried out for each of the 3 classification problems (IPF/non-IPF, NSIP/non-NSIP, DIP/non-DIP). In each of the experiments the dataset was randomly partitioned into three approximately equal parts, two of which were used as the training set, and the third one as the testing set. By rotating the subset taken in the role of the test set, in fact each experiment consisted of 3 tests, i.e. a total of 60 experiments were carried out for each of the three classification problems. The average accuracy of these 1800 experiments (i.e. five methods applied 60 times to original and reduced datasets of three problems) measured on the test sets is shown in Table 2.

It can be seen from Table 2 that by removing the two outliers, the accuracy of every single classification method was improved for each of the 3 models.

Table 2

Classification Accuracies Before/After Elimination of Outliers

	Dataset	MultilayerPerceptron	SimpleLogistic	SMO	IB1	J48	Average change in accuracy
NSIP/non-NSIP	Original	72.00%	72.19%	75.35%	66.96%	71.32%	+6.25%
	Reduced	79.26%	78.33%	79.63%	75.37%	76.48%	
IPF/non-IPF	Original	80.27%	81.78%	81.25%	70.66%	82.74%	+2.08%
	Reduced	82.87%	84.07%	82.04%	72.87%	85.28%	
DIP/non-DIP	Original	84.07%	87.81%	88.23%	84.58%	85.97%	+3.29%
	Reduced	89.07%	88.80%	90.37%	88.15%	90.74%	

In conclusion, the suspicions generated by the weakness of the coverage with patterns of two of the observations, lead to the identification of these two patients as outliers, and eventually to medical explanations of the inappropriateness of maintaining them in the dataset. The "cleaned" dataset obtained by eliminating these two outliers was shown to allow a substantial improvement in the accuracy of all the tested classification methods.

4 Support Sets

4.1 Set covering formulation

Although the dataset involves 13 variables, some of them may be redundant. Following the terminology of LAD ([6], [7], [8]) we shall call an irredundant set of *variables* or *attributes* or *features* a *support set* of the dataset, if projecting on this subset the 13 dimensional vector representing the patients, there will be no overlap between the 3 different types of IIPs.

The determination of a minimum size support set was formulated as a set covering problem. The basic idea of the set covering formulation of this problem consists in the simple observation that a subset S is a support set if and only if the projections on S of the positive and the negative observations in the dataset are disjoint.

In order to illustrate this reduction we shall identify a minimum size subset of the variables in the dataset which are capable of distinguishing IPF observations from non-IPF observations. We shall assume that the three numerical variables x_{11}, x_{12}, x_{13} have been

"*binarized*", i.e. each of them had been replaced by one or several 0-1 variables, as proposed in ([5], [6]). The binarized variables are associated to so-called *cut-points*. For instance, there are 2 cut-points (5.5 and 6.5) associated to the numerical variable x_{11}, and the corresponding binary variables $x_{11}^{5.5}$ and $x_{11}^{6.5}$ are then defined in the following way:

$$x_{11}^{5.5}=1 \text{ if } x_{11}\geq 5.5, \text{ and } x_{11}^{5.5}=0 \text{ if } x_{11}<5.5,$$

$$x_{11}^{6.5}=1 \text{ if } x_{11}\geq 6.5, \text{ and } x_{11}^{6.5}=0 \text{ if } x_{11}<6.5.$$

Similarly, two cut-points (7.5, 8.5) are introduced for x_{12}, along with two associated binary variables. The variable x_{13} is binarized using four 0-1 variables associated to the cut-points 0.5, 1, 1.05 and 1.2.

Using the original 10 binary variables along with the 8 binarized variables (which replace the numerical variables x_{11}, x_{12}, x_{13}), we shall now represent the observations as 18 dimensional binary vectors $(x_1,\ldots,x_{10}, x_{11}^{5.5}, x_{11}^{6.5}, x_{12}^{7.5}, x_{12}^{8.5}, x_{13}^{0.5}, \ldots, x_{13}^{1.2})$. For example, the positive (i.e. IPF) observation s008 = (0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 2, 9, 0.22) will become in this way the binary vector b008 = (0, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0). Similarly the negative (i.e. non-IPF) observation s006 = (0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 5, 2, 2.5) becomes the binary vector b006 = (0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1).

Clearly, the positive binarized observation b008 and the negative binarized observation b006 differ only in the following 8 components: x_2, x_3, $x_{12}^{7.5}, x_{12}^{8.5}, x_{13}^{0.5}, \ldots, x_{13}^{1.2}$. It follows that any support set S must include at least one of these variables, since otherwise the projections on S of the positive observation b008 and the negative observation b006 could not be distinguished. Therefore, if we denote by $(s_1,\ldots,s_{10}, s_{11}^{5.5}, s_{11}^{6.5}, s_{12}^{7.5}, s_{12}^{8.5}, s_{13}^{0.5}, \ldots, s_{13}^{1.2})$ the characteristic vector of S, one of the necessary conditions for S to be a support set is

$$s_2 + s_3 + s_{12}^{7.5} + s_{12}^{8.5} + s_{13}^{0.5} + \ldots + s_{13}^{1.2} \geq 1.$$

A similar inequality can be written for every pair consisting of a positive (IPF) and a negative (non-IPF) observation in the binarized dataset. The 34*22=748 pairs of positive-negative observations define the constraints of a set covering problem for finding a minimum size support set. Since our dataset consists of a rather limited number of observations, in order to increase the accuracy of the models to be built on the support sets obtained in this way, we have further strengthened the above set covering-type constraints, by replacing the 1 on their right-hand side, by 3 (the choice of 3 is based on empirical considerations, the basic idea being simply to sharpen the requirements of separating positive and negative observations).

Clearly the objective function of this set covering type problem is simply the sum

$$s_1+\ldots+s_{10} + s_{11}^{5.5} + s_{11}^{6.5} + s_{12}^{7.5} + s_{12}^{8.5} + s_{13}^{0.5} + \ldots + s_{13}^{1.2}.$$

4.2 Three minimum support sets

By solving this problem we found that the binary variables x_3, x_4, x_9, x_{10} are redundant, and that a minimum size support set (using the original binary and numerical variables) consists of the attributes 1, 2, 5, 6, 7, 8, 11, 12 and 13.

In a similar way we can see that a minimum support set distinguishing DIP observations from non-DIP ones consists of the 6 original attributes: 1, 2, 3, 5, 12 and 13,
while a minimum support set distinguishing NSIP patients from non-NSIP ones consists of the 8 original attributes: 1, 2, 5, 6, 7, 8, 11 and 12.

4.3 Accuracy of classification on minimum support sets

It is important to point out that the elimination of redundant variables does not reduce the accuracy of classification. In order to demonstrate the qualities of the minimum support sets obtained for the IPF/non-IPF, DIP/non-DIP and NSIP/non-NSIP problems we have carried out 20 three-folding classification experiments on these 3 problems using 5 different classification methods from the WEKA package ([*http: //www.cs.waikato.ac.nz/~ml/weka/index.html*]); these experiments used first the original 13 variables, and after that the support sets of 9, 6, and 8 variables respectively, obtained above for these 3 problems. The results of these experiments are presented in Table 3.

Table 3
Classification Accuracies on all Original Variables and on Support Sets

	Support Set	MultilayerPerceptron	SimpleLogistic	SMO	IB1	J48	Average change in accuracy
NSIP/non-NSIP	Original	79.26%	78.33%	79.63%	75.37%	76.48%	+2.48%
	Reduced	85.19%	78.80%	79.35%	81.57%	76.57%	
IPF/non-IPF	Original	82.87%	84.07%	82.04%	72.87%	85.28%	+0.98%
	Reduced	85.28%	83.33%	81.20%	76.39%	85.83%	
DIP/non-DIP	Original	89.07%	88.80%	90.37%	88.15%	90.74%	+1.89%
	Reduced	91.76%	90.28%	89.44%	92.87%	92.22%	

In conclusion we can see from Table 3 that the elimination of those features which were identified as redundant does not only maintain the accuracy of classification, but actually increases it in each of the three models.

5 Patterns and Models

Using the support sets developed in the previous section, we shall apply now the LAD methodology to this dataset for generating patterns and classification models. It turns out that in spite of the very small size of this dataset, some surprisingly strong patterns can be identified in it. For example in the IPF/non-IPF model, 14 (i.e. 70%) of the 20 non-IPF patients satisfy the simple pattern "GG2/RET $\geq$ 1.2"; moreover none of the 34 IPF patients satisfy this condition. While this simple pattern involves a single variable, other more complex patterns exist and are capable of explaining the IPF or non-IPF character of large groups of patients. For instance, the negative pattern

"RET $\leq$ 8 and GG2/RET > 1"

is satisfied by 70% of the non-IPF patients, and by none of the IPF patients. As an example of a positive pattern, we mention

"HC = 1, HPL = 0 and GG2/RET $\leq$ 1.2";

24 (i.e. 70.6%) of the 34 IPF patients satisfy all the 3 constraints of this pattern, and none of the non-IPF patients satisfy simultaneously these 3 conditions.

While the above patterns can distinguish large groups of patients having a certain type of IIP from those of other types of IIP, larger collections of patterns constructed by LAD can classify collectively the entire set of 54 observations in the dataset. We shall first illustrate the way the classification works by considering the problem of distinguishing IPF and non-IPF patients.

We present in Table 4 a model consisting of 20 positive and 20 negative patterns allowing the accurate classification of IPF/non-IPF patients. Note that the equality of the numbers of positive and negative patterns in this model is a simple coincidence.

The way in which the model allows the classification of a new observation is the following. First, if an observation satisfies all the conditions describing some positive (negative) patterns, but does not satisfy all the conditions describing any one of the negative (positive) pattern, then the observation is classified as positive (negative); this classification is shown in the tables as "1" (respectively, "0"). Second, if an observation does not satisfy all the defining conditions of any positive or negative pattern, then it remains "unclassified"; this is shown in the tables as "?". Third, if an observation satisfies all the defining conditions of some positive and also of some negative patterns in the model, then a weighting process is applied to decide on the appropriate classification; the process of finding weights for such classification is described in ([3]).

Beside the IPF/non-IPF model discussed above, we have also constructed a model to distinguish the 14 NSIP patients from the 40 non-NSIP patients, and another model to distinguish the 6 DIP patients from the 48 non-DIP patients. The NSIP/non-NSIP model is built on the support set of 8 attributes described in the previous section, and includes 16 positive and 4 negative patterns. The DIP/non-DIP model is built on the support set of 6 attributes described in the previous section, and includes 7 positive and 15 negative patterns.

The combination of the three models allows the drawing of additional conclusions. For example, if the results of the three classifications are 0, 0 and ? respectively, and one knows that

each patient is exactly of one type of IIP, one can conclude that the "?" in the classification of the third condition can be replaced by "1".

Table 4

IPF/non-IPF model

Pattern	attr.1	attr.2	attr.5	attr.6	attr.7	attr8	attr.11	attr.12	attr.13	Pos Prevalence	Neg Prevalence
	IIT	HC	BRVX	PL	HPL	SL	GG2	RET	GG2/RET		
P1		1			0				≤1.2	70.6%	0
P2		1			0			≥8		47.1%	0
P3	1	1					≥4		≤1.2	47.1%	0
P4		1			0	0		≥6		47.1%	0
P5	1	1			0			≥6		47.1%	0
P6	1	1					≥4	≥8		41.2%	0
P7		1	0						≤0.5	41.2%	0
P8		1	0					≤8	≤1.2	41.2%	0
P9	1	1							>0.5, ≤1.2	38.2%	0
P10				1					≤1.2	32.4%	0
P11	1	1						≥8	>0.5	32.4%	0
P12	1	1						≥9		29.4%	0
P13		1	0				≤3			29.4%	0
P14		1					≥4	≤8	≤1.2	26.5%	0
P15					0			≥8	≤0.5	26.5%	0
P16			0	1				≥6		26.5%	0
P17	0							≤8	≤1.2	20.6%	0
P18						1		≥8		20.6%	0
P19				1			≤3			20.6%	0
P20				1	0					17.6%	0
N1								≤8	>1	0	70.0%
N2									>1.2	0	70.0%
N3		0				0	≥4			0	50.0%
N4								≤5		0	50.0%
N5		0				0			>0.5	0	50.0%
N6		0		0	0					0	45.0%
N7		0		0				≤7		0	45.0%
N8		0			0				>0.5	0	40.0%
N9		0			0		≥4			0	40.0%
N10					0				>1	0	40.0%
N11		0		0		0				0	40.0%
N12	0								>1	0	35.0%
N13	1	0					≥4	≤7		0	30.0%
N14					1	0		≤8	>0.5	0	30.0%
N15					1	0	≥4	≤8		0	30.0%
N16				0	1	0		≤8		0	20.0%
N17						1			>1	0	15.0%
N18	0				1	0	≥4			0	15.0%
N19			1		1	0		≤8		0	15.0%
N20	0				1	0			>0.5	0	15.0%

The results of the classification of the 54 patients given by the three models, along with the conclusions derived from the knowledge of all the three classifications are presented in Table 5. The accuracy of this classification is 100%.

Table 5

Observations	Given Classification	Classification by LAD Models			
		IPF/nonIPF	NSIP/nonNSIP	DIP/nonDIP	Conclusion
s001	DIP	0	?	1	DIP
s002	DIP	0	0	1	DIP
s004	DIP	0	0	1	DIP
s005	DIP	0	?	1	DIP
s006	DIP	0	?	1	DIP
s007	DIP	0	0	1	DIP
s008	IPF	1	0	0	IPF
s009	IPF	1	?	0	IPF
s010	IPF	1	?	0	IPF
s011	IPF	1	0	0	IPF
s012	IPF	1	0	0	IPF
s013	IPF	1	0	0	IPF
s014	IPF	1	0	0	IPF
s015	IPF	1	0	0	IPF
s016	IPF	1	?	0	IPF
s017	IPF	1	0	0	IPF
s018	IPF	1	0	0	IPF
s019	IPF	1	0	0	IPF
s020	IPF	1	0	0	IPF
s021	IPF	1	?	0	IPF
s022	IPF	1	0	0	IPF
s023	IPF	1	0	0	IPF
s024	IPF	1	0	0	IPF
s025	IPF	1	0	0	IPF
s026	IPF	1	0	0	IPF
s027	IPF	1	0	0	IPF
s028	IPF	1	0	0	IPF
s029	IPF	1	0	0	IPF
s030	IPF	1	0	0	IPF
s031	IPF	1	0	0	IPF
s032	IPF	1	0	0	IPF
s033	IPF	1	0	0	IPF
s034	IPF	1	0	0	IPF
s035	IPF	1	0	0	IPF
s036	IPF	1	0	0	IPF
s037	IPF	1	0	0	IPF
s038	IPF	1	0	0	IPF
s039	IPF	1	0	0	IPF
s040	IPF	1	0	0	IPF
s041	IPF	1	0	0	IPF
s042	NSIP	0	1	0	NSIP
s043	NSIP	0	1	?	NSIP
s044	NSIP	0	1	0	NSIP
s045	NSIP	0	1	0	NSIP
s047	NSIP	0	1	0	NSIP
s048	NSIP	0	1	?	NSIP
s049	NSIP	0	1	0	NSIP
s050	NSIP	0	1	0	NSIP
s051	NSIP	0	1	?	NSIP
s052	NSIP	0	1	0	NSIP
s053	NSIP	0	1	0	NSIP
s054	NSIP	0	1	0	NSIP
s055	NSIP	0	1	0	NSIP
s056	NSIP	0	1	0	NSIP

6 Validation

It has been shown in the previous section (Table 5) that the accuracy of classifying by LAD the 54 patients is of 100%. It should be added however that this result represents only the correctness of the proposed classification model when the entire dataset is used both as a training set, and as a test set. In order to establish the reliability of these classifications they have to be validated. Because of the very limited size of the dataset (in particular because of the availability of data for only 6 DIP patients and only 14 NSIP patients) the traditional partitioning of the dataset into a training and a test set would produce extremely small subsets, and therefore highly unreliable conclusions. In view of this fact, we shall test the accuracy of the LAD classification by cross-validation, using the so-called "jackknife" or "leave-one-out" method. As an example, the cross-validation of the classification results for the IPF/non-IPF model will be presented in the next section.

The basic idea of the "leave-one-out" method is very simple. One of the observations is temporarily removed from the dataset, a classification method is "learned" from the set of all the remaining observations, and it is applied then to classify the extracted observation. This procedure is then repeated separately for every one of the observations in the dataset. For example in the case of the IPF/non-IPF model we have to apply this procedure 54 times.

Table 6 shows the results of the "leave-one-out" procedure applied to this model. The table includes the results of directly applying leave-one-out experiments to the 3 models (IPF/non-IPF, NSIP/non-NSIP, DIP/non-DIP), as well as the resulting combined classifications. The combined classifications are then used to derive the final conclusion about the IPF/non-IPF character of each observation; the correctness of the conclusion (compared with the given classification) is presented in the last column of Table 6 ("evaluation").

It can be seen that out of 54 observations, 44 are classified correctly, there are 6 errors (the IPF patients s009, s010 and s021 are classified as non-IPF, and the non-IPF patients s042, s047 and s053 are classified as IPF), two patients (s007 and s052) are unclassified, and for two other patients (s016 and s055) the classifications ("IPF or NSIP") are imprecise.

If one considers every unclassified and every imprecisely classified patient as an error, the accuracy of the classification in the leave-one-out experiment is 81.48%. However, if we use the formula established in [3] for accuracy, this turns out to be 85.80%.

In view of the very small size of the dataset, the results of the leave-one-out tests can be viewed as extremely encouraging.

Table 6

Validation by Leave-One-Out of IPF/non-IPF Classification

Obs.	Given Classification	Classification by Leave-One-Out			Derived Classification	Conclusion	
		IPF/nonIPF	NSIP/nonNSIP	DIP/nonDIP		IPF/nonIPF	Evaluation
s001	DIP	0	?	1	DIP	0	correct
s002	DIP	0	0	1	DIP	0	correct
s004	DIP	0	0	1	DIP	0	correct
s005	DIP	0	1	1	DIP or NSIP	0	correct
s006	DIP	0	1	1	DIP or NSIP	0	correct
s007	DIP	0	0	0	?	?	unclassified
s008	IPF	1	0	0	IPF	1	correct
s009	IPF	0	?	0	NSIP	0	error
s010	IPF	0	1	0	NSIP	0	error
s011	IPF	1	0	0	IPF	1	correct
s012	IPF	1	0	0	IPF	1	correct
s013	IPF	1	0	0	IPF	1	correct
s014	IPF	1	0	0	IPF	1	correct
s015	IPF	1	0	0	IPF	1	correct
s016	IPF	1	1	0	IPF or NSIP	?	imprecise
s017	IPF	1	0	0	IPF	1	correct
s018	IPF	1	0	0	IPF	1	correct
s019	IPF	1	0	0	IPF	1	correct
s020	IPF	1	0	0	IPF	1	correct
s021	IPF	0	1	0	NSIP	0	error
s022	IPF	1	0	0	IPF	1	correct
s023	IPF	1	0	0	IPF	1	correct
s024	IPF	1	0	0	IPF	1	correct
s025	IPF	1	0	0	IPF	1	correct
s026	IPF	1	0	0	IPF	1	correct
s027	IPF	1	0	0	IPF	1	correct
s028	IPF	1	0	0	IPF	1	correct
s029	IPF	1	0	0	IPF	1	correct
s030	IPF	1	0	0	IPF	1	correct
s031	IPF	1	0	0	IPF	1	correct
s032	IPF	1	0	0	IPF	1	correct
s033	IPF	1	0	0	IPF	1	correct
s034	IPF	?	0	0	IPF	1	correct
s035	IPF	1	0	0	IPF	1	correct
s036	IPF	1	0	0	IPF	1	correct
s037	IPF	1	0	0	IPF	1	correct
s038	IPF	1	0	0	IPF	1	correct
s039	IPF	1	0	0	IPF	1	correct
s040	IPF	1	0	0	IPF	1	correct
s041	IPF	1	0	0	IPF	1	correct
s042	NSIP	1	0	0	IPF	1	error
s043	NSIP	0	1	?	NSIP	0	correct
s044	NSIP	0	1	0	NSIP	0	correct
s045	NSIP	0	1	0	NSIP	0	correct
s047	NSIP	1	?	0	IPF	1	error
s048	NSIP	0	?	1	DIP	0	correct
s049	NSIP	0	1	0	NSIP	0	correct
s050	NSIP	0	1	0	NSIP	0	correct
s051	NSIP	0	1	?	NSIP	0	correct
s052	NSIP	0	0	0	?	?	unclassified
s053	NSIP	1	0	0	IPF	1	error
s054	NSIP	0	1	0	NSIP	0	correct
s055	NSIP	1	1	0	NSIP or IPF	?	imprecise
s056	NSIP	0	1	0	NSIP	0	correct

7 Attribute Analysis

7.1 Importance of attributes

A simple measure of the importance of an attribute is the frequency of its inclusion in the patterns appearing in the model. For example, attribute 1 (IIT) appears in 11 (i.e. in 27.5%) of the 40 patterns of the IPF/non-IPF model in Table 4.

The frequencies of all the 13 attributes in the models are shown in Table 7 for the 3 LAD models considered, along with the averages of these 3 indicators.

Table 7

Frequencies of Attributes in Models

Attributes	IPF/non-IPF	NSIP/non-NSIP	DIP/non-DIP	Average
IIT	0.275	0.25	0.343	0.289
HC	0.525	0.813	0.357	0.565
TB	0	0	0.238	0.079
GG1	0	0	0	0.000
BRVX	0.125	0.219	0.381	0.242
PL	0.2	0.094	0	0.098
HPL	0.4	0.375	0	0.258
SL	0.3	0.156	0	0.152
AC	0	0	0	0.000
N	0	0	0	0.000
GG2	0.25	0.688	0	0.313
RET	0.5	0.5	0.376	0.459
GG2/RET	0.475	0	0.662	0.379

Two of the most important conclusions which can be seen in this table indicate that:

- the most influential attributes are honeycombing (HC), reticulation score (RET), ground-glass attenuation/reticulation score (GG2/RET), and ground-glass attenuation score (GG2);
- the attributes ground-glass attenuation (GG1), airspace consolidation (AC) and nodules (N) have no influence on the classification.

7.2 Promoting and Blocking Attributes

We shall illustrate the promoting or blocking nature of some attributes on the IPF/non-IPF model shown in Table 4. It can be seen in the table that every positive pattern which includes a condition on HC (honeycombing) requires that HC=1. Conversely, every negative pattern which includes a condition on HC requires that HC=0. This means that if a patient is known to be a non-IPF case with HC=1, and all the attributes of another patient have identical values except for HC which is 0, then this second patient is certainly not an IPF case. This type of monotonicity means simply that HC is a “promoter” of IPF. It is easy to see that the attribute PL(polygonal lines) has a similar property.

On the other hand, the attribute BRVX (peri-bronchovascular thickening) appears to have a converse property. Indeed, every positive pattern which includes this attribute requires that BRVX=0, while the only negative pattern (N19) which includes it requires that BRVX=1. Therefore if a patient's BRVX would change from 1 to 0, the patient's condition would not change from IPF to non-IPF (assuming again that none of the other attributes change their values). Similarly to the previous case, this type of monotonicity means simply that BRVX is a "blocker" of IPF.

In this way the IPF/non-IPF model allows the identification of two promoters and of one blocker. None of the other attributes in the support set appear to be promoters or blockers.

A similar analysis of the DIP/non-DIP model shows that intralobular interstitial thickening (IIT) and traction bronchiectasis (TB) are blockers of DIP. Also, the analysis of the NSIP/non-NSIP model shows that peri-bronchovascular thickening (BRVX) is a promoter of NSIP, while honeycombing (HC), polygonal lines (PL) and septal lines (SL) are blockers of NSIP.

To conclude, we show in Table 8 the promoters and blockers which have been identified for the three forms of idiopathic interstitial pneumonias.

Table 8

	Idiopathic Pulmonary Fibrosis	**Desquamative Interstitial Pneumonia**	**Non Specific Interstitial Pneumonia**
honeycombing	promoter		blocker
polygonal lines	promoter		blocker
peri-bronchovascular thickening	blocker		promoter
intralobular interstitial thickening		blocker	
traction bronchiectasis		blocker	
septal lines			blocker

8 Conclusions

It has been shown that it is possible to use a computational technique (LAD) for analyzing CT data for distinguishing with high accuracy different entities (IPF, NSIP and DIP) of idiopathic interstitial pneumonias (IIPs). This is particularly important for NSIP which is as yet poorly defined. It was also shown that the patterns developed by LAD techniques provide additional information about outliers, redundant features, the relative significance of the attributes, and allow to identify promoters and blockers of various forms of IIPs. These encouraging results will 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.

9 References

[1] G. Alexe, S. Alexe, P.L. Hammer, L. Liotta, E. Petricoin, M. Reiss, Logical Analysis of Proteomic Ovarian Cancer Dataset, *Proteomics*, 4, 2004, 766-783.

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

[3] S. Alexe and P.L. Hammer, Pattern-Based Discriminants in the Logical Analysis of Data, *In this Volume*.

[4] American Thoracic Society / European Respiratory Society International Multidisciplinary Consensus. Classification of the Idiopathic Interstitial Pneumonias. *Amer J Respir Crit Care Med*, 165, 2002, 277-304.

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

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

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

[8] 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.

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