34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Characterizing Non-Linear Dependencies Among Pairs of Clinical Variables and Imaging Data

Jesus J. Caban[1,2], Ulas Bagci[2], Alem Mehari[3], Shoaib Alam[3],
Joseph R. Fontana[3], Gregory J. Kato[3], Daniel J. Mollura[2]

Abstract— **Advances in computer-aided diagnosis (CAD) systems have shown the benefits of using computer-based techniques to obtain quantitative image measurements of the extent of a particular disease. Such measurements provide more accurate information that can be used to better study the associations between anatomical changes and clinical findings. Unfortunately, even with the use of quantitative image features, the correlations between anatomical changes and clinical findings are often not apparent and definite conclusions are difficult to reach. This paper uses nonparametric exploration techniques to demonstrate that even when the associations between two-variables seems weak, advanced properties of the associations can be studied and used to better understand the relationships between individual measurements.**

This paper uses quantitative imaging findings and clinical measurements of 85 patients with pulmonary fibrosis to demonstrate the advantages of non-linear dependency analysis. Results show that even when the correlation coefficients between imaging and clinical findings seem small, statistical measurements such as the maximum asymmetry score (MAS) and maximum edge value (MEV) can be used to better understand the hidden associations between the variables.

I. INTRODUCTION

During the last decade, many CAD systems have been adopted as clinical tools from which physicians and radiologists can obtain quantitative information about the progression of a particular disease [1]. In chest CT, CAD systems have been used to automatically identify and measure pulmonary infection such as influenza [2], pulmonary fibrosis [3], Tree-in-Bud nodularity [4], and many other diseases [1].

Although image analysis has proven to be crucial in the detection and monitoring of numerous diseases and CAD systems have shown to be effective at assisting radiologists during the interpretation and decision-making process, in general the associations between imaging and physiological findings is still not well understood. The main challenge that arises when correlating image and clinical findings is the complexity and nonlinearity aspects of the relationship.

During the last two decades multiple techniques to analyze the relationship between individual variables in large datasets have been proposed. To a large extent, current data analysis techniques often used in biomedical applications have not kept abreast of these advanced and powerful approaches. Some of the most popular techniques are mutual information estimators [5], maximal correlations [7], principal curve-based methods [6], distance correlation [8], and more recently maximal information coefficient (MIC) [9].

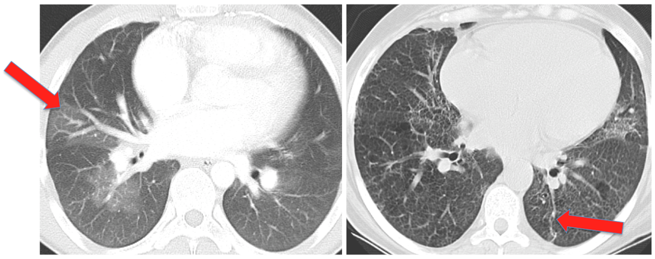

Fig. 1. (left) Ground glass opacity is an ill-defined increase attenuation of the lungs with preserved visibility of lung architecture. (right) Linear opacity (arrow) can represent fibrotic scarring.

We hypothesize that in many cases there are relevant associations between imaging patterns and clinical findings, however the analysis must go beyond linear correlations techniques.

To explore and better understand the non-linearity properties between imaging and clinical data, a dataset of 85 patients with pulmonary fibrosis was used from protocols approved by the Institutional Review Board of the National Heart, Lung and Blood Institute (ClinicalTrials.gov identifiers NCT00023296, NCT00081523, or NCT00352430). First, a computer-based application was designed to identify and quantify ground glass opacities (GGO) and linear interstitial (fibrosis). GGO is an ill-defined increased attenuation of the lungs caused by different lung diseases as illustrated in Figure 1(left). Linear opacity is illustrated in Fig. 1(right). The first step of our research was to design and develop a texture-based CAD system to quantitatively recognize pulmonary CT features [2].

In addition to the chest CT studies, each patient also had a pulmonary function test (PFT) done to obtain different physiological measurements of the lungs. PFT is used to measure airflow and the volume of air the lungs take and release during inhalation and exhalation.

There are still no clear understandings of the associations between clinical measurements such as PFT values and patterns observed by radiologists in imaging studies. The major limitation has been the weak correlation that is often found when comparing individual variables.

[1]JJ. Caban is with the NICoE, Naval Medical Center E: jesus.caban@nih.gov

[2]U. Bagci and D.J. Mollura are with the Center for Infectious Disease Imaging, Radiology and Imaging Science Department, National Institutes of Health. E: ulas.bagci@nih.gov, molluradj@cc.nih.gov

[3] GJ. Kato, A. Alam, A. Mehari, and JR Fontana are with National Heart, Lung and Blood Institute, NIH.

This paper presents how nonparametric techniques can be used to explore the non-linearity properties of the data and to estimate additional statistical measurements for determining the relevance of different clinical features.

II. BACKGROUND

The relationship between two random variables x_i and x_j can be explored by estimating the correlation and dependencies between them. Linear correlations including the Pearson product-moment correlation and regression have been the most widely accepted techniques to study the associations between individual clinical variables [10].

The continuing interest in effective ways for findings associations between a pair of variables have motivated the design and implementation of other correlation models such as mutual information estimators [5], maximal correlations [7], principal curve-based methods [6], and distance correlation [8]. Recently a new maximal information-based heuristic technique was introduced named the *Maximal Information Coefficient* (MIC) [9].

A. Maximal Information Coefficient (MIC)

MIC is a novel measure of dependence that captures linear and non-linear associations between pair of variables. MIC basically constructs a grid with various sizes and finds the largest mutual information obtained from the pairwise data. Nevertheless, MIC is not an estimate of mutual information, but a rank order statistic helping to understand the underlying complexity of the data. As I denotes the mutual information and G denotes the particular grid, MIC of a set D of pairwise data with a sample size n and grid size less than $B(n)$ is given by

$$MIC(D) = \max_{xy<B(n)}\{M(D)_{x,y}\} \tag{1}$$

where

$$M(D)_{x,y} = \frac{I^*(D,x,y)}{log\,min(x,y)}, \tag{2}$$

and $I^*(D,x,y) = \max\ I(D|_G)$ for different distributions of grids G such that $B(n)$ is the maximal grid size and for practical reasons it is set to $n^{0.6}$ (see [9] supplementary notes for details).

MIC has three key properties that were used to explore the non-linear properties of the data and to develop a feature selection procedure. These features are *maximum asymmetry score* (MAS), *maximum edge value* (MEV), and *minimum cell number* (MCN). The following subsections, we briefly describe these properties and their use in the proposed feature selection framework.

Maximum Asymmetry Score: MAS is a measure of non-monotonicity. Monotonic functions are those functions that follow a particular order, thus monotonic properties are very useful for differentiating individual functions. In general, non-monotonicity is not a desirable property due to lack of consistency in the functional analysis. MAS, in the context of MIC computation, is a measure of non-monotonicity that basically captures the deviation of a function from monotonicity. Based on all these definitions, MAS can be recalled as non-consistency as well. In short, the low values of MAS indicate differentiable, consistent, and well-posed relationship among the data. By following Reshef et al [9], MAS can be defined as

$$MAS(D) = \max_{xy<B}|M(D)_{x,y} - M(D)_{y,x}| \tag{3}$$

where $M(D)$ is characteristic matrix of a set D of pairwise variable data.

Maximum Edge Value (MEV): MEV is defined as a closeness to being a function and measures the degree to which the dataset appears to be sampled from a continuous function [9]. For a given set D, it is defined as

$$MEV(D) = \max_{xy<B}\{M(D)_{x,y} : x \text{ or } y = 2\} \tag{4}$$

MEV is ranged from 0 to 1 such that the large values of MEV indicate well behaved functions.

Minimum Cell Number (MCN): MCN is known as complexity measure which simply counts the number of cells required to reach the MIC score. While well-defined and monotone functions require less number of cells, non-monotone and parametrically poorly defined functions require large number of cells to reach MIC. MCN can be defined simply as

$$MCN(D,\epsilon) = \min_{xy<B} log(xy) \tag{5}$$

where $M(D)_{x,y} \geq (1-\epsilon)MIC(D)$ and ϵ is a robustness term and depends on the MIC such that for noiseless case ϵ is considered to be 0.

III. APPROACH

In this paper we postulate that the correlation measurement between two variables must be jointly considered with other properties of the data in order to effectively determine the associations between imaging and clinical findings. In particular, correlation measurements and advanced statistical properties of the data such as the MAS to determine the monotonicity of the data, the MEV to understand the closeness to a function, and MCN to know the complexity of the function must be combined and used to determine the relevance of the associations.

One specific advantage of using advanced statistical properties of the data is to better understand the meaning of the weak correlations as those often seen between imaging and clinical findings. In addition, the combination of multiple advanced statistical properties can be used to create a more elaborated feature selection technique that jointly consider multiple statistical aspects of the data.

A. Feature Selection Based on MIC and Its Key Properties

We propose to use MIC and its key properties as a feature selection algorithm in order to explore non-linear associations among clinical and imaging findings where conventional linear correlation analysis falls short. Since MIC and MEV are proportional, and MAS and MCN are inversely proportional to the pairwise relationship of data as described in previous section, we create a novel metric by combining the MIC and its key features as

$$\theta_i = \prod_i r_i^{MIC} \frac{r_i^{MEV}}{(r_i^{MAS} r_i^{MCN})} \qquad (6)$$

where i represents particular pairwise relationship (i.e., PFT versus GGO).

Data	MIC	MAS	MEV	MCN	θ
Linear plot	1	0	1	2	inf
Parabola	1	0.69	1	2.56	0.56
Two Lines	0.79	0.16	0.70	6.91	0.50
Circular	0.71	0.03	0.32	6.87	1.06
Circular + Noise	0.46	0.19	0.22	6.98	0.07

TABLE I
MAXIMAL INFORMATION COEFFICIENT (MIC) AND SOME OF ITS ADVANCED NON-LINEAR STATISTICAL PROPERTIES WERE ESTIMATED FOR A SET OF SYNTHETIC FUNCTIONS TO BETTER UNDERSTAND COMPLEX ASSOCIATIONS IN CLINICAL DATA.

Table I shows a set of well defined functions with their corresponding MIC, MAS, MEV, MCN, and θ properties. From the table we can see that linearly correlated data will have high MIC and MEV values. Some of the advantage of using nonparametric exploration techniques is that even the associations in non-linear functions such as a parabola or a circle can be captured. By using the MIC, the key properties of MIC, and θ, we can approximate the association between two random variables to a given function as shown Table I.

IV. RESULTS

To show the limitations of linear correlation techniques when analyzing the associations between image and clinical variables, linear regression was used to obtain the correlation coefficients R and R^2 between abnormal imaging patterns of CT (GGO and fibrosis) against clinical measurements obtained from the PFT test. Table II shows some of the results when applying linear correlation to analyze the associations between GGO and Fibrosis with PFT. From the results we can see that the most significant correlation when considering GGO obtained an R value of 0.394 and an R^2 value of 0.155. The results show that there is a weak, but significant correlation between the amount of GGO and the Expiratory Reserve Volume (ERV). A similar correlation appears when analyzing fibrosis.

Only the Expiratory Reserve Volume (ERV), timed forced expiratory volumes (FEV1), total lung capacity (TLC), and the vital capacity (VC) were weakly correlated with GGO and linear opacity (fibrosis). Table II also shows that similar correlations were found when analyzing the associations between PFT parameters and fibrosis. However, from the results we can see that in general PFT parameters are correlated more closely with GGO than with fibrosis.

When analyzing the top PFT variables in Table II using the advanced non-linear properties of the data, we can better understand the type of associations that exist between PFT and imaging findings. Figure 2(top) shows that the MIC values of the first four variables range from 0.240 to 0.312. That shows that the association between the data does not follow a particular linear pattern, instead a more circular or torus-type of relationship as demonstrated by the synthetic results shown in Table I. This is, some data points reside inside the circles, some around the circumference of the circle, and some outside the shape.

	Ground Glass Opacity (GGO)		Fibrosis	
Name	R	R^2	R	R^2
ERV	0.394	0.155	0.347	0.120
FEV1	0.330	0.109	0.314	0.099
TLC	0.311	0.097	0.239	0.057
VC	0.322	0.104	0.329	0.108
FVC	0.282	0.080	0.298	0.089
FRC	0.228	0.052	0.143	0.020
RV	0.218	0.048	0.082	0.007
IC	0.118	0.014	0.171	0.029

TABLE II
TABLE WITH SAMPLE ASSOCIATIONS BETWEEN PTF AND IMAGING FEATURES.

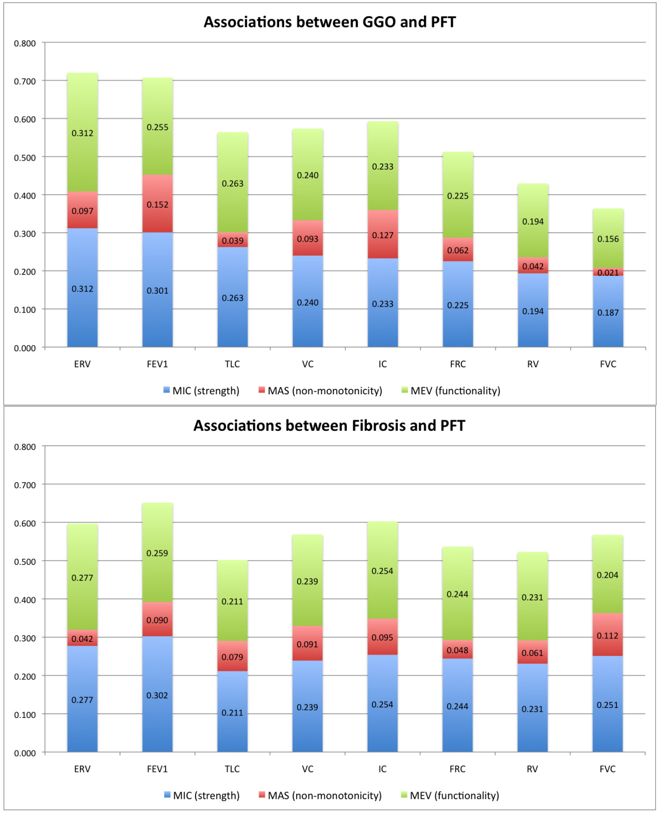

Fig. 2. Analysis of the associations between PFT and imaging features such as GGO and Fibrosis. For each clinical measurement under consideration, the MIC, MAS, MEV properties of the data were estimated.

Furthermore, when we analyze the monotonicity of the associations, we can see that most of them are between 0.039 and 0.152. This is a very important finding because it means that the relationships between PFT and imaging findings follow a particular order. The monotonicity pattern means that as the amount of GGO increases or decreases, there is a clear pattern that the data points get closer or farther away from the center of the circle. Finally, when we analyze the MEV results we can see that the values range between 0.240 and 0.312. By looking at Table I, we can infer that there is a non-linear circular function that can be used to capture the associations between those PFT parameters and imaging features. Similar results can be found when comparing PFT measures against the volume of fibrosis as illustrated in Fig. 2 (bottom).

We found that only the ERV, FEV1, TLC, and VC were found to be weakly correlated with GGO and fibrosis. Similar correlations were found when analyzing the associations between PFT and fibrosis. However, in general it seems that PFT is more correlated with GGO than with fibrosis.

Ground Glass Opacities (GGO)	
Name	Relevance
TLC	0.497
FVC	0.363
ERV	0.281
RV	0.248
FRC	0.228
VC	0.172
FEV1	0.132
IC	0.118

TABLE III
TABLE SHOWING THE RANKING OF THE PFT FEATURES WHEN USING OUR FEATURE SELECTION TECHNIQUE.

By using our features selection technique explained in Section III, we can determine the relevance of a set of given clinical measurements. When testing our feature selection technique with PTF, we found that the relevance of the measurements slightly changed more into accordance with what is expected from the physiological point of view. For instance, the total lung capacity (TLC) might be expected to be more highly correlated with the volume of air a patient can inhale and exhale. Thus, demonstrating and taking into account the non-linear properties of the data in conjunction with correlation properties opens new ways to look at the associations between multi-modal clinical data. Table III shows some of the results when using our feature selection technique. Note that any reliable relevance can be encoded as a functional form (i.e., high MEV values) which can give insights into determining unique ordered pairs, eventually leading to bio-marker identification.

V. Discussion and Conclusion

It is proved that MIC is more powerful when data is noiseless (MEV is high and/or MAS is low), however, is still not addressed how to deal with complex data including certain amount of noise (i.e., due to measurement errors, sensitive clinical variables, etc). To cope with this statistical power problem of MIC, we are aiming to bring into attention flexible grids which allows the use of various size and shape parameter to encapsulate noisy relationships better. Furthermore, in order to solve this issue, an information fusion system can be adapted by taking into account the rank of statistical powers of various different approaches (such as distance correlation) other than MIC.

In our implementation of feature selection based on MIC and its key features, we assumed that the importance of each key feature contributed equally to the construction of θ. Indeed, depending on the nature of the problem, certain weights can be approximated to feature selection as

$$\theta_i^{new} = \prod_i (r_i^{MIC})^{\alpha} \frac{(r_i^{MEV})^{\beta}}{(r_i^{MAS})^{\gamma}(r_i^{MCN})^{\omega}} \qquad (7)$$

such that $\alpha + \beta + \gamma + \omega = 1$. However, this is outside the scope of current paper and will be evaluated separately as an extension of the current formulation.

In this study, we analyzed the relationships of the clinical and imaging variables based on novel data association measures. Based on this, we developed a simple yet efficient feature selection system to build a bridge from system biology to medical imaging. We demonstrated that finding the relationships between spaces spanned by clinical and imaging variables are beyond the horizons of simple linear correlations, and more sophisticated methods are needed to ensure that the complexity of the relationship is handled. As an extension of this work, we aim to build a reliable probabilistic decision mechanism based on the proposed feature selection algorithm together with a near-optimal classification algorithm to predict possible bio-markers for the particular imaging abnormalities pertaining to lung diseases.

ACKNOWLEDGMENT

This research was supported in part by the Center for Infectious Disease Imaging, Intramural Research Program, at the National Institutes of Health.

References

[1] Bagci U et al., ”Computer-assisted detection of infectious lung diseases: a review.”, Comput Med Imaging Graph. 2012 Jan;36(1):72-84

[2] Yao J. et al., ”Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification”, Academic Radiology (2011) V. 18(3)

[3] Kim HG et al, ”A computer-aided diagnosis system for quantitative scoring of extent of lung fibrosis in scleroderma patients” Clin Exp Rheumatol. 2010 Sep-Oct;28

[4] Bagci U et al., ”Learning Shape and Texture Characteristics of CT Tree-in-Bud Opacities for CAD Systems”, In Proc. of MICCAI, 2011

[5] Moon Y et al.,”Estimation of using kernel density estimators” Physical Review, vol. 52, no. 3, pp. 2318–2321, 1995

[6] Delicado P., ”Measuring non-linear dependence for two random variables distribute along a curve”, Statistics and Computing, vol. 19, no. 3, pp. 255-269, 2009

[7] Yu Y., ”On the maximal correlation coefficient”, Statistics and Probability Letters, Volume 78(9), pp. 10721075, 2008

[8] G. Szekely et al., ”Brownian distance covariance”, Applied Statistics, vol. 3, pp. 1236–1265, 2009.

[9] Reshef D et al, ”Detecting Novel Associations in Large Data Sets”. Science 334 (6062): pp. 1518–1524 2011

[10] Huber P, ”Robust Statistics”, Wiley Press, 2004