ON-LINE DEPOSITORY
Camiciottoli G, Bigazzi F, Paoletti M, Cestelli L, Pistolesi M.
Data analysis

We started the CT data analysis investigating the possible relationships between %LAA-950 and AWT-Pi10. We evaluated the presence of a linear association between the two variables by the Pearson's correlation index [1]. A *r* value of 0.25 ($p<0.05$) was found, indicating a poor linear association. The distribution of CT data is plotted in Figure 1A of the paper. Visual exploration of this data set reveals the absence of structured relationships between the two CT parameters. Accordingly, it appears quite difficult to define a parametric procedure to classify patients according to these direct measures. For this reason, we merged the information of the original measurements (%LAA-950 and AWT-Pi10) by principal component analysis. This permit to obtain two novel numeric variables, that we called CT1 and CT2, capable to classify patients considering the two alterations at the same time. CT1 is proportional to the difference between the original CT variables (%LAA-950 and AWT-Pi10), hence representing the prevailing mechanism of airflow limitation (parenchymal destruction or conductive airway obstruction). CT2 is proportional to the sum of the two original variables, hence representing the overall COPD severity as resulting from both parenchymal destruction and conductive airway obstruction.

The novel CT indexes can be introduced as a linear combination of the primary CT variables according to the following equations:

$$CT1 = (\%LLA\text{-}950 - AWT\text{-}Pi10)/\sqrt{2} \quad (1)$$
$$CT2 = (\%LLA\text{-}950 + AWT\text{-}Pi10)/\sqrt{2} \quad (2)$$

As a direct consequence of CT1 and CT2 indexes definition we can assert:
1) increasing CT1 values means that the parenchymal destruction (%LLA-950 contribution) is predominant on airway obstruction (AWT-Pi10 contribution), and vice-versa.
2) increasing CT2 values means that the overall severity is increased (%LLA-950 and AWT-Pi10 contributions).

From a mathematical point of view these transformations correspond to project the original standardized data onto the Principal Components plane as briefly explained in the following paragraphs (the $\sqrt{2}$ quantity is a normalization factor deriving from the math argumentation.)

Principal components analysis is a mathematical method often used to reduce the dimensionality of the data while retaining most of the variation in the data set. The reduction is performed by identifying those directions called principal components along which the variation in the data is maximal [2]. The goal of this method is to concentrate the information about the differences between samples into a small number of dimensions. In particular a set of n-dimensional vector samples x = {x1, x2, x3 …, xm} should be transformed into another set y = {y1, y2, …, ym} of the same dimensionality, but having the property that most of their information (in terms of variance) content is stored in the first few dimensions. This will allow us to reduce the data set to a smaller number of dimensions with low information loss.

Formally, if ***x*** is a random vector with covariance matrix Σ and mean μ, then we define the *principal component transformation (or Karhunen-Loewe transformation)* [3] as follows:

$$x \to y = \Gamma'(x - \mu)$$

where Γ is orthogonal, $\Gamma'\Sigma\Gamma = \Lambda$ is a diagonal matrix and the eigenvalues are ordered as follow:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4 \ldots \geq \lambda_p \geq 0 \ldots .$$

The i-*th principal component* of ***x*** is defined as the *i*-th element of the vector ***y***:

$$y_i = \gamma_i (x - \mu)$$

where γ_i is the *i*-th column of the matrix Γ.

In general, there are as many components as variables. However, because of the way they are calculated, it is usually possible to consider only a few of the principal components which together explain *most* of the information contained in the primary data matrix. Many criteria have been suggested for deciding how many principal components to retain [3]. Principal components analysis is a complex theme [3] and a specific knowledge of linear algebra is required to understand the mathematical details underlying principal components analysis-related techniques. However, we introduced here only the basic concepts to introduce a simple geometrical interpretation of equations 1 and 2. Figure 1 (on-line

depository) shows the distribution of the standardized CT data (each dot represents a patient 's AWT-Pi10 value plotted against its corresponding %LAA-950) together with the two principal components. Principal components analysis identifies the two directions (red and blue lines) along which the data have the largest spread. In particular, it is possible to see that the first component (red) is the direction along which the samples show the largest variation. The second component (blue) is the secondary direction, uncorrelated and orthogonal to the first component, along which the samples show the largest variation (obviously in our example we have only two components since the CT measures are represented by two variables).

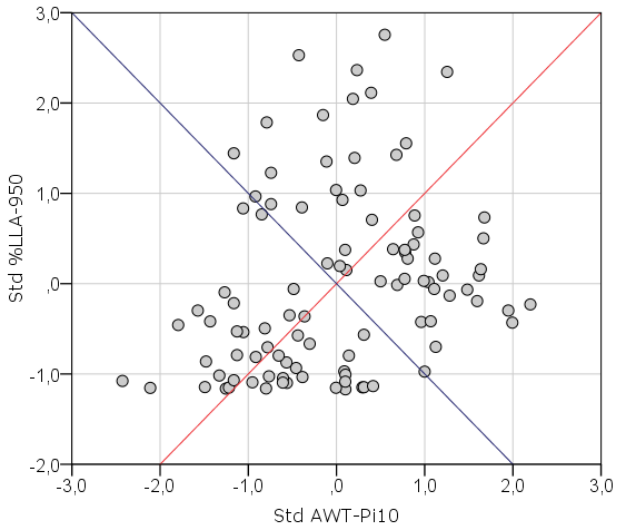

Figure1 (on-line depository)

From an algebra point of view, if data are standardized such that each variable is centered to zero average, the principal components are the normalized eigenvectors of the covariance matrix and ordered according to how much of the variation present in the data they contain. Each component can then be interpreted as the direction, uncorrelated to previous components, which maximizes the variance of the individual samples when projected onto the component itself. Since in our example we have only two standardized variables, the two principal components directions are simply represented by the two lines explained by the following equations:

Std %LLA-950=Std AWT-Pi10 (red line)

and

Std %LLA-950= -Std AWT-Pi10 (blue line)

The calculation proposed in (1) and (2) correspond to transform the original CT data points by projecting each couple of CT measures in the principal components space. In facts the new coordinates are simply the distances of each point from the two lines representing the two principal components. Figure 1B of the paper represents the transformed data after applying the equations 1 and 2. Each patient is represented in the new graph by the two indexes CT1, CT2 instead of the original CT measures.

CT indexes modeling and COPD patient classification

The two CT indexes are candidate to represent a valid diagnostic decision making support to classify patients according to their specific phenotype and overall severity (see the published paper for details). To make accessible this kind of classification to/for patients with no CT data available, we trained and validated two predictive models to estimate these scores considering clinical and functional variables as independent variables. In particular we identified a couple of simple multivariate linear models to classify prospectively COPD patients on the basis of their *estimated* CT parameters instead of the true features extracted from expensive imaging techniques. For each CT index, a subset of optimal predictors was defined using a forward stepwise process (*F-statistics* was evaluate to enter or remove parameters [4]). The prediction performances of the models were then evaluated through a ten-fold cross-validation process [5]. The obtained models, regression coefficients and cross validation *R-shrinkage* are summarized in Table 4 of the paper.

The estimated *CT1* and *CT2* scores can be written as:

CT1= -0.018xDLco + 0.011xTLC – 0.58xSputum + 0.324

CT2=-0.03xFEV1/VC + 0.013xFRC + 0.775xSputum-0.575

Sputum is a *boolean* variable defined as *"1"* if chronic purulent sputum is present, "0*"* otherwise. Other parameters are numerical variables derived directly by PFT.

Using the two CT indexes as linear classifiers for prospective COPD patients, they could implement the following decision rule: - Decide “phenotype A” if CT1 > 0 or “phenotype B” if CT1 < 0 and “major severity” if CT2>0 or “minor severity” if CT2<0 (see the paper for the clinical interpretation of the CT indexes). If CT1, CT2 = 0, the corresponding patient can ordinarily be assigned to either class, but for simplicity we shall leave the assignment undefined here. Higher values of the CT scores will take to a strongest membership to the corresponding class.

Following this interpretation the equations CT1 = 0 and CT2 = 0 define the *decision surfaces* that separates patients assigned to the different classes. Since we trained two linear models, these surfaces are simple hyper planes depending by the predictors and the scores CT1 and CT2 represent the distances of each patient from the two decision surfaces.

Let us write the two boundary-equations *CT1=0* and *CT2=0* as:

0.02 x *DLco + 0.56* x *Purulent sputum = 0.01* x *TLC + 0.48*

and

0.03 x $FEV_1/VC = 0.01$ x *FRC + 0.57* x *Purulent sputum + 0.01*

The interpretation of the two decision planes is now straightforward; in fact one patient will lay on the decision surface (*CT1 equilibrium*) if the weighted TLC value is able to contrast the mixed effects of Dlco and Sputum characteristics. The result is that *Dlco* and *Purulent sputum* are two aligned co-factors in phenotype determination in opposition with *TLC*. Similar considerations can be done for CT2 where FEV_1/VC seems to be in opposition to *Purulent sputum* and *FRC* mixed contributions in determining the COPD severity index CT2.

Obviously, the reliability of a linear classifier is deeply influenced by the robustness of the coefficients identification process and by the capacity of the learning set to be representative of the whole population. In this study, the decision boundaries have been defined following a data-driven approach, training the two CT models on a learning dataset derived directly

from true CT images. In opposition to other well-known aprioristic COPD indexing techniques where the various scores are assigned arbitrarily, each patient is classified here on the basis of a set of clinical co-factors. In addition, the contribution of each factor to the CT1 and CT2 scores is properly weighted through coefficients obtained after an optimization approach. In this manner we have realized a well calibrated and bias-free classification system for prospective COPD patients.

However, to correctly classify the various prospective patterns, we strongly suggest to consider the obtained indexes as fuzzy scores and to consider an ambiguous region near the zero to take in account of the wide spectrum of this pathology and the variability of the measures.

As an applicative example, after model validation, CT indexes were estimated in 373 COPD patients who did not undergo CT, using their clinical and PFT data as predictors. Fig. 2 of the paper shows their distribution plotted into the PCs space, each patients is represented by a couple of estimated CT indexes. Also for these patients we investigated the relationships between the predicted scores and their clinical parameters calculating the Pearson correlation coefficient between CT indexes and PFTs variables. ANOVA was performed to analyze means variability of CT indexes compared with some categorical parameters. Table E1 reports the distribution of the categorical variables in the four subsets of patients identified in Figure 2.

TABLE E1. Percent prevalence of categorical variables in the four subsets of the 373 patients of the *testing set* subdivided according to the values of predicted CT1 and CT2.

	A n=143	B n=77	C n=80	D n=73
	CT1<0 CT2<0	CT1>0 CT2<0	CT1<0 CT2>0	CT1>0 CT2>0
Cough 0 = absent	17,5	36,4	1,3	24,7
Cough 1 = occasional	23,1	33,8	20	32,9
Cough 2 = chronic	59,4	29,9	78,8	42,5
Sputum 0 = absent/occasional	24,5	57,1	3,8	47,9
Sputum 1 = chronic non purulent	27,3	41,6	1,3	17,8

Sputum 2 = chronic purulent	48,3	1,3	95	34,2
mMRC 0 = none	11,2	11,7	3,8	1,4
mMRC 1 = slight	22,4	19,5	17,5	8,2
mMRC 2 = moderate	39,9	31,2	46,3	31,5
mMRC 3 = severe	23,1	31,2	27,5	39,7
mMRC 4 = very severe	3,5	6,5	5	19,2

mMRC: modified Medical Research Council dyspnoea score; Chronic cough and chronic purulent sputum prevail ($p<.01$) in A and C; Severe and very severe mMRC dyspnoea scores prevails in B and D.

References (on-line depository)

1. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. *The American Statistician* 1988; 42: 59–66.

2. Larose DT. Data Mining Methods and Models. New York: Wiley; 2006.

3. Jolliffe IT. Principal Component Analysis, Second Edition. New York: Springer; 2002.

4. Larose DT. Statistical approaches to estimation and prediction. In: Larose DT, ed. Discovering knowledge in data: an introduction to data mining. 1st edn. New York John Wiley & Sons 2005; pp 83-88.

5. Picard R, Cook D. "Cross-Validation of Regression Models". *J Am Statist Assoc* 1984; 79: 575–583.