Online supplementary data – Arthritis and Rheumatism manuscript reference: ar-07-0691

A paradigm of diagnostic criteria for polyarteritis nodosa: analysis of a series of 949 vasculitides

Corneliu Henegar, Christian Pagnoux, Xavier Puéchal, Jean-Daniel Zucker, Véronique Le Guern, Mona Saba, Denis Bagnères, Olivier Meyer, and Loïc Guillevin for the French Vasculitis Study Group (FVSG)

A FORMAL PRESENTATION OF SEVERAL KEY ANALYTICAL FEATURES OF THE STUDY

The purpose of this additional document is to provide a short formal presentation of selected key methodological characteristics of this study. As indicated in the main manuscript, the analytical design used in this study comprised two distinct stages. The first stage was to select a minimum set of low redundant positive and/or negative PAN predictive items, among those exhibiting the highest individual accuracy in distinguishing PAN from other systemic vasculitides in the FVSG patient sample. The selection of this set relied on clinical judgment supported by a combination of uni- and multivariate statistical analysis of clinical and paraclinical items used to describe patient characteristics in the FVSG database. During the second stage, the selected set of criteria was evaluated through an unsupervised computer-simulation procedure, designed to reproduce the case-based aspect of the clinical diagnostic reasoning. Both analytical stages were compared to the 1990 ACR classification criteria considered to be the most reliable reference to date. The aim of the computer-simulation procedure was two-fold: first, to test the dependence of the PAN-predictive performances of these sets on the prevalence of individual vasculitides in the analyzed patient samples; and, second, to evaluate their

robustness to the random noise, which may affect distributions of clinical and paraclinical parameters in real conditions.

Analysis of capabilities of available clinical and paraclinical items to predict PAN

The univariate analysis performed to assess the strength of individual associations between PAN diagnosis and available clinical or paraclinical features relied on a normalized, pairwise, mutual information (MI) measure, which is a well-established approach for quantifying positive or negative mutual dependence between two variables (1). MI computation depends on the notion of entropy of a random variable derived from Shannon's theory of information. Briefly, for a discrete random variable X whose probability distribution is $p(X = x_i),\ i = 1,\ldots,N_x$, where N_x is the number of all possible values of X, its entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{N_x} p(X = x_i)\log_2 p(X = x_i) \tag{1}$$

Based on equation (1), the pairwise MI of two random variables X, Y can be computed as:

$$MI(X,\ Y) = H(X) + H(Y) - H(X,\ Y) \tag{2}$$

where $H(X, Y)$ denotes the joint entropy of the two variables. Intuitively, MI measures how much knowing the value of one of these variables reduces our uncertainty about the other. If X and Y are independent, then X contains no information about Y and vice versa, so their MI is zero. If X and Y are identical then all information conveyed by X is shared with Y, therefore the MI is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y. The normalized MI, $\overline{MI}(X, Y)$, is a relative measure derived from (2) which reduces the influence of the magnitudes of individual entropies:

$$\overline{MI}(X, Y) = \frac{MI(X, Y)}{\max\{H(X), H(Y)\}} \tag{3}$$

To establish a minimal set of low-redundant PAN-predictive criteria we conducted an exploratory multivariate analysis that relied on available parameters to build a logistic-regression model through a forward stepwise inclusion approach. The inclusion procedure was directed by clinical judgment and the PAN-predictive value of individual items, as determined by the univariate analysis, and was reiterated until the logistic-regression model reached saturation (e.g. until a chi-square goodness-of-fit test performed after each step could no longer detect any significant improvement by further inclusion of additional items).

Computer simulation of PAN-predictive abilities of FVSG and 1990 ACR criteria

During a second analytical stage, computer simulations were run to evaluate the PAN-predictive abilities of the two sets of criteria under various conditions simulated through artificially generated vasculitis patient data. The Boolean aspect of the presence of clinical and paraclinical features in vasculitis patients suggested the possibility of relying on a model of aggregated dependent Bernoulli trials to represent the real joint distributions of the clinical and paraclinical parameters specific to each vasculitis. We considered two distinct approaches to quantify and express marginal distributions, and dependencies between individual parameters.

The first approach relied on the Bahadur–Lazarsfeld theoretical framework (2), which computes a complete representation of the joint distribution $p(x)$ of a set of n correlated Bernoulli trials (e.g. corresponding to *n* clinical or paraclinical items) through an expansion of a binomial law:

$$p(x) = p_{[1]} f(x) \tag{4}$$

where $p_{[1]}$ is the product of marginal distributions of *n* Bernoulli trials $x_1,...,x_i,...,x_n$,

$$p_{[1]}(x_1,...,x_n) \equiv \prod_{i=1}^{n} p_i^{x_i}(1-p_i)^{1-x_i} \tag{5}$$

and $f(x)$ is a correction factor in terms of n marginal probabilities p_i and of $2^n - n - 1$ correlation parameters expressing dependencies between Bernoulli trials (e.g. including all correlations r from the 2nd to the *n*th order) as:

$$f(x) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<k} r_{ijk} z_i z_j z_k + ... + r_{12...n} z_1 z_2 ... z_n \tag{6}$$

where $z_i = (x_i - p_i) \Big/ \sqrt{p_i(1-p_i)}$. Despite its good theoretical accuracy, an obvious drawback of the Bahadur–Lazarsfeld expansion is its requirement of a high number of dependency parameters, which challenges the computational tractability of the model, even for a moderate number of trials. On the other hand, tentative approximations of the Bahadur–Lazarsfeld model by truncation (e.g. by considering as input parameters only marginal probabilities and second-order correlations) were shown to be less robust than the original model, as they could result in some abnormal outcome results (e.g. negative or >1) (3). All these considerations suggested the potential usefulness of a recently proposed theoretical solution, which relies on the maximum entropy principle to optimize the inference of missing parameters of a truncated Bahadur–Lazarsfeld expansion (3). Moreover, this approach has another relevant feature, as it allows verification of the consistency of input parameters (e.g. marginal probabilities and second-order correlations), and its restoration if necessary by projecting the input parameters into the feasible domain, thereby achieving a highly precise reproduction of the clinical and paraclinical characteristics of real vasculitis patients in artificially generated data. The required marginal

distributions and the second-order correlations between clinical and paraclinical items were computed from the patient sample extracted from the FVSG database.

The second approach used to generate artificial patient data relied on a maximum-spanning tree (MST) dependence-modeling technique, which approximates dependencies between individual parameters by arbitrarily limiting them to those expected to have the most impact on the results (4, 5). The effect of this limitation would be increased random noise in the generated data.

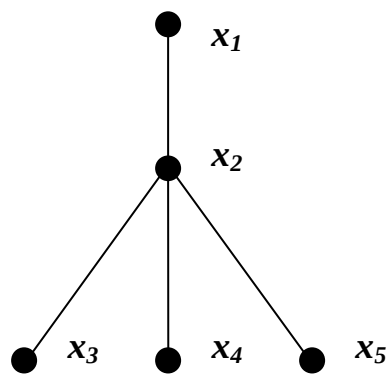

Figure 1. An example of a maximum-spanning tree (MST) representation of the strongest dependency relationships among five clinical and paraclinical items $x_1, \ldots, x_5$.

The principle of the MST dependence-modeling technique resides in estimating the joint distribution of item presence by relying on an MST representation of their strongest interdependences. Thus, given two nodes, *i* and *j*, of a spanning-tree representation (e.g. corresponding to two clinical and paraclinical items), such that the ith node is directly and immediately above the jth node, an MST may be defined as that maximizing the sum

$$\sum_{i,j} MI(node_i, node_j) \tag{7}$$

where $MI(node_i, node_j)$ represents the expected mutual information provided by node *i* about node *j*, computed from equations (1) and (2). For example, if we consider the MST representation of the dependencies between five parameters (1 by 1), $x_1,...,x_5$ depicted in figure 1, it is possible to compute the joint probability of the combined occurrence of x_1, x_3, x_5, based on the information that one node provides about another neighboring node, as:

$$p(x_1, x_3, x_5) = p(x_1 = 1) \times p(x_3 = 1 \mid x_2 = 0) \times p(x_5 = 1 \mid x_2 = 0) \times p(x_2 = 0 \mid x_1 = 1) \times p(x_4 = 0 \mid x_2 = 0) \quad (8)$$

The marginal distributions of clinical and paraclinical items and the pairwise MI coefficients required by the MST approach were computed from the patient sample extracted from the FVSG database.

The artificial patient data thus generated was further used in a computer simulation to evaluate the usefulness the two sets of criteria when used to screen potential vasculitis patients for positive PAN diagnosis. To achieve this goal, we relied on a combination of an unsupervised hierarchical clustering approach, used to group artificially generated cases based on the similarities of their clinical and paraclinical profiles, and a supervised labeling procedure, which assigns to each resulting cluster of similar cases the true label of the cases that form the majority of its content.

The clustering approach starts by grouping the two most similar cases together to form a first cluster, and then reiterates the agglomerative procedure until all cases are collected in a single cluster, thereby generating a new partition of cases into clusters at each iterative step. The choice of the optimal partition of clusters is a fundamental issue in unsupervised learning. A popular solution to this problem is to simplify it by finding the partition that provides the best

trade-off between the homogeneity of the clusters and their isolation on the partition (6). Although there is no best approach to fit all situations, the computation of the Silhouette index was shown to be a simple yet robust strategy for the prediction of optimal clustering partitions (7). This method assigns to each sample *i* of a given cluster C_j $(j = 1, ..., c)$, a quality measure $s(i)$ $(i = 1, ..., m)$, known as the Silhouette width. This measure is an indicator of confidence in the membership of the ith sample in cluster C_j, and is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{9}$$

where $a(i)$ is the average dissimilarity between the ith sample and all the samples included in C_j, and $b(i)$ is the minimum average dissimilarity between the ith sample and all of the samples clustered in C_k $(k = 1, ..., c;\ k \neq j)$. From equation (9) it follows that $-1 \leq s(i) \leq 1$. When $s(i)$ is close to 1, it is considered that the ith sample has been well clustered. When $s(i)$ is close to 0, it is suggested that the ith sample could also be assigned to the nearest neighboring cluster. If $s(i)$ is close to −1, it can be argued that this sample has been misclassified. Thus, for a given cluster C_j, it is possible to compute a cluster Silhouette S_j that characterizes the heterogeneity and isolation properties of C_j:

$$S_j = \frac{1}{m} \sum_{i=1}^{m} s(i) \tag{10}$$

where m is the number of samples in C_j. It has been shown that for any partition $P = \{C_1 \cup ... C_j \cup ... C_c\}$, a global Silhouette value, GS_P, can be used as an effective validity index for partition *P*.

$$GS_P = \frac{1}{c}\sum_{j=1}^{c} S_j \tag{11}$$

After identifying the optimal partition and the supervised labeling of its clusters, the predictive abilities of the two sets of criteria were evaluated by computing estimations of sensitivity, specificity, and of positive- and negative-predictive values, from the contingency table reflecting the attribution of cases to each vasculitis.

References

1. Yao YY. Information-theoretic measures for knowledge discovery and data mining. In: Karmeshu, editor. Entropy Measures, Maximum Entropy Principle and Emerging Applications. 1st ed. Berlin, Germany: Springer; 2003. p. 115-136.
2. Bahadur RR. A representation of the joint distribution of responses to *n* dichotomous items. In: Solomon H, editor. Studies in Item Analysis and Prediction. 1st ed. Stanford, CA: Stanford University Press; 1961. p. 158-168.
3. Van Der Geest PAG. The binomial distribution with dependent Bernoulli trials. Journal of Statistical Computation and Simulation 2005;75:141-154.
4. Yu CT, Buckley C, Lam K, Salton G. A generalized term dependence model in information retrieval. Information Technology: Research and Development 1983;2:129-154.
5. Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 1968;14:462-467.
6. Xu R, Wunsch Dn. Survey of clustering algorithms. IEEE Trans Neural Netw 2005;16:645-78.

7. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987;20:53-65.