Discriminant analyses of peanut allergy severity scores

Olivier Collignon, Jean-Marie Monnez, Pierre Vallois, F. Codreanu, J.-M. Renaudin, Gisèle Kanny, Marie Brulliard, Bernard Bihain, Sandrine Jacquenet, Denise-Anne Moneret-Vautrin

Discriminant analyses of peanut allergy severity scores

O.Collignon[1,2], J.-M.Monnez[2], P.Vallois[2], F.Codreanu[3], J.-M.Renaudin[1], G.Kanny[3], M.Brulliard[1], B.E.Bihain[1], S.Jacquenet[1], D.Moneret-Vautrin[3]

[1] *Genclis SAS, 15 rue du Bois de la Champelle, 54500 Vandoeuvre-lès-Nancy, France*;
[2] *Institut Elie Cartan, UMR 7502, Nancy Université, CNRS, INRIA, BP239, 54506, Vandoeuvre-lès-Nancy, France* ;
[3] *Centre Hospitalier Universitaire, Service d'allergologie, 29 av. Mar De Lattre de Tassigny, 54000 Nancy, France*

Abstract

Peanut allergy is one of the most prevalent food allergies. The possibility of a lethal accidental exposure and the persistence of the disease make it a public health problem. Evaluating the intensity of symptoms is accomplished with a double blind placebo controlled food challenge (DBPCFC), which scores the severity of reactions and measures the dose of peanut that elicits the first reaction. Since DBPCFC can result in life-threatening responses, we propose an alternate procedure with the long term goal of replacing invasive allergy tests. Discriminant analyses of DBPCFC score, the eliciting dose and the first accidental exposure score were performed in 76 allergic patients using 6 immunoassays and 28 skin prick tests. A Multiple Factorial Analysis was performed to assign equal weights to both groups of variables and predictive models were built by cross-validation with LDA, k-NN, CART, penalized SVM, stepwise logistic regression and AdaBoost methods. We developed an algorithm for simultaneously clustering eliciting dose values and selecting discriminant variables. Our main conclusion is that antibody measurements offer information on the allergy severity, especially those directed against *rAra-h1* and *rAra-h3*. Further independent validation of these results and the use of new predictors will help extend this study to clinical practices.

Keywords : *discriminant analysis, peanut allergy, DBPCFC, Multiple Factorial Analysis, classification, variable selection*

1 Introduction

An allergy is an abnormal reaction of the immune system towards foreign substances (*allergens*) that are normally harmless. Peanut allergies in particular affect more than 0.5% of the entire French population, and its increasing prevalence and potentially severe clinical reactions make it a public health problem. It is also the most lethal food allergy [4]. Following a strict avoidance diet is currently the only effective treatment that minimises potentially lethal accidents.

Diagnosing and scoring peanut allergies is currently performed with a *double blind placebo controlled food challenge (DBPCFC)* [3]. Patients are given increasing peanut doses until the first clinical reaction appears. Those showing specific allergy symptoms are declared allergic, and a particular avoidance treatment is then initiated. DBPCFC is also used to judge the severity of an established peanut allergy by determining the cumulative dose that triggers the first reaction, known as the *eliciting dose* [in *milligrams (mg)*]. However, these tests require patient hospitalisation in specialised centers and can potentially result in life-threatening reactions from patients with severe allergies. The DBPCFC is also a costly and time consuming test to conduct.

The severity of peanut allergies is usually scored using the following scale [1] :

- **Score 1:** Mild symptoms among : abdominal pains that spontaneously resolve under 30 minutes and/or rhinocunjunctivitis and/or urticaria < 10 papulas and/or a rash (eczema onset);
- **Score 2:** One moderate symptom among : abdominal pain requiring treatment or generalized urticaria or non-laryngeal angioedema or cough or fall of Peak Expiratory Flow between 15 and 20%;
- **Score 3:** Two moderate symptoms in the preceding list;
- **Score 4:** Three moderate symptoms in the preceding list or laryngeal oedema or hypotension or asthma requiring treatment;
- **Score 5:** Any symptom requiring hospitalisation in intensive care.

For an already diagnosed allergy, it would be much more advantageous to predict the severity of the reaction from accidental exposure by using a blood sample or cutaneous test. This would replace the DBPCFC test with a simple statistical tool that can still evaluate potential risk without exposing the patient to a life-threatening allergic situation. Such a diagnostic method would be a major advance in food allergies and be beneficial to both patients and clinicians.

The first objective of this paper was to select a set of discriminant variables which can offer useful information about the severity of peanut allergy. These variables could provide biologists and allergists a better understanding of the mechanisms inducing allergic reactions. Moreover this will allow to avoid measuring useless variables in further studies.

The second goal of this study was to predict the DBPCFC score, the eliciting dose and the first accidental exposure score, evaluated *a posteriori* with the patient's medical record according to the same scale as the DBPCFC score. The first accidental exposure score would then reveal the "real" severity of the allergy. Compare this to using the DBPCFC, which only offers a minimal view of the severity since the procedure is terminated once the first symptoms appear.

2 Experimental Procedure and Data

A clinical study was performed using 76 allergic patients with ages from 3 to 18 years. Tables 1 and 2 describe the frequencies observed for DBPCFC score, the first accidental exposure score and the eliciting dose. Note that only 47 out of the 76 patients experienced a first accident. The remaining 29 patients were diagnosed during an allergy check-up and subsequently confirmed by DBPCFC, thus avoiding further accidents. Patients were homogeneously distributed in age and sex across severity scores.

Thirty-four variables were measured to reveal the presence of *Immunoglobulins of type E (IgE)* antibodies. These are proteins produced by the immune system that can elicit allergic reactions [20]. Each antibody is specific to an allergen, *i.e.*, it is coded to identify a particular protein for elimination. We measured the levels of *IgE* for the proteins of interest with the goal of building a predictive model of allergy severity. The variables used to test for *IgE* were measured either by immunoassays or by Skin Prick Tests (SPTs).

2.1 Immunoassays

Immunoassays are biochemical tests that quantify the level of antibodies in a blood sample (in *kilo-units per liter*). We performed six immunoassays aimed at measuring the following: the *total IgE*, the *specific IgE to peanut (f13)*, and the *specific IgE to recombinant (r)Ara-h1, rAra-h2, rAra-h3, rAra-h8*, which are *IgE* especially directed against peanut recombinant major allergens [1].

2.2 Skin Prick Tests (SPTs)

SPTs are used to detect an immunological sensitivity to a particular substance. They show the functional aspect of cellular *IgE*, which are linked to mast cells releasing chemical mediators that elicit symptoms [20]. A small dose of allergen is applied under the skin by pricking with a needle, and the diameter of the resulting wheal is measured in millimeters. We also measured the diameter of prick-tests to codeine as a positive control showing the basal reactivity of the skin. The ratio of the two diameters is used to measure the allergen reaction.

We performed prick-tests for 28 allergens divided into three families:

1. **11 nuts**: *almond, Brazil nut, cashew nut, chestnut, hazelnut, peanut, pecan nut, pine nut, pistachio, Queensland nut, walnut*, which are often related to peanut allergies by cross-reactivity;

2. **7 legumes**: *broad bean, chickpea, dried bean, green pea, lentil, lupine flour, soybean*, since peanuts are legume;

3. **10 aeroallergens**: *12 grass pollens, Alternaria, ash, birch, cat epithelia, dog epithelia, Dpte (Dermatophagoïdes pteronyssinus), mugwort, ribwort, rape seed*, which are the common clinical allergens.

Immunoassays and SPTs were measured immediately before DBPCFC.

3 Statistical approach

All computations were performed using the SAS Enterprise Guide 4.1.0.471 ® or R 2.7.0 [21].

3.1 Design of the study

We first performed a Principal Component Analysis (PCA) to gain an overview of the data.

To solve our problems, discriminant analyses of DBPCFC score, first accident score and eliciting dose were performed by using several classifiers. Two studies were performed for each measure of severity by treating it as a four-class variable, and then as a two-class variable.

For DBPCFC and first accidental exposure, 4 classes were built by considering the score groups $\{1\}, \{2\}, \{3\}$ and $\{4, 5\}$ because of the low frequency of score 5. For the two-class discrimination, groups were formed by scores of either $\{1, 2, 3\}$ or $\{4, 5\}$, as recommended by clinicians.

Since eliciting dose values are fixed by levels by the clinician [1], its measure is not a continuous variable and cannot be predicted by a regression analysis. Moreover although eliciting dose could be considered as a class variable, a discriminant analysis cannot be directly performed because of numerous categories with low frequencie (Table 2). A first solution consisted in converting eliciting dose into a four or two-class variable by searching the best discriminated classification computed with all available variables. A second solution will be proposed in Section 3.3.

Careful variable selection appeared especially as a major point of the analysis. Therefore three different statistical approaches were proposed for each measure of severity :

3.1.1 Direct application of the classifiers

The performances of several classification rules were first compared without preselecting variables. Linear Discriminant Analysis (LDA), k-Nearest Neighbors ($k-$NN), Classification And Regression Trees (CART) [10] and AdaBoost with CART [5] were performed using all the 34 available variables as predictors. $k-$NN were performed for $k \in \{1, \ldots, 5\}$ and the number of nearest neighbors giving the best results was kept.

3.1.2 Simultaneous variable selection and classification

Since in supervised learning keeping noisy predictors can increase the misclassification error, two methods that simulteanously perform variable selection and classification were also used : stepwise logistic regression [11] and penalized SVM [2].

3.1.3 Specific variable selection scheme prior to classification

DBPCFC and first accident scores

As explained earlier, the determination of a set of predictors to keep is one of the main points of the study. Thus a variable selection scheme independent from classification was also developed. Variables were retained in the model if either the corresponding p-value of the Kruskal-Wallis test [6] was smaller than 0.10, or if the variable was selected by the stepwise Wilks' lambda (Λ) criterion [13]. The Λ statistic was computed at each step with all variables already present in the model, whereas the $F - to - enter$ statistic and corresponding p-value measure the discriminant power of a variable added to the preceding ones. For the latter, the maximal $F - to - enter$ p-values used as entry and removal criteria were set by default to 0.15, as recommended by [7].

The nonparametric Kruskal-Wallis test was preferred to ANOVA, because variables were not always normally distributed in the classes induced by the scores. Note that the Wilks' lambda selection is based on the hypotheses of multi-normality of the variables vector distribution and equality of the within-class covariance matrices. In 1975, Lachenbruch [14] asserted that the $F-$test is robust to small deviations of these hypotheses. We therefore decided to use the Wilks' lambda selection even though variables were not always normally distributed in the classes induced by the measures of severity.

This variable selection scheme is a compromise between the assessment of variable marginal importance and the detection of a discriminant subset of predictors. This will allow first to provide biologists and allergists a list of informative variables in regards to the mechanisms involved in allergic reactions, and second to avoid keeping noisy variables that could degrade the performance of the learning algorithms.

Biologists and allergists are particularly interested in using immunoassays as markers of the severity of peanut allergy because *Specific IgE to rAra-h1, rAra-h2, rAra-h3* are yet known to be very useful in practice to detect peanut allergic patients [1]. Moreover immunoassays are precise and reliable measures contrary to SPTs. But the number of discriminant SPTs can far exceed the number of selected immunoassays, which could possibly smear out the signal brought by immunoassays. As we wanted nevertheless to keep the information provided by SPTs, a Multiple Factorial Analysis (MFA) [8] was performed to equalize the influence of both groups of selected variables, which enabled the use of factors as new predictors of severity.

Thus we performed two discriminant analyses :

1. by using directly the discriminant variables as predictors for classification with the methods introduced in Section 3.1.1,

2. by computing MFA factors of the discriminant variables and selecting a limited number of discriminant factors by the same selection process as the one used with raw variables, before performing the classification rules.

The overall specific statistical approach is summarized in Figure 1.

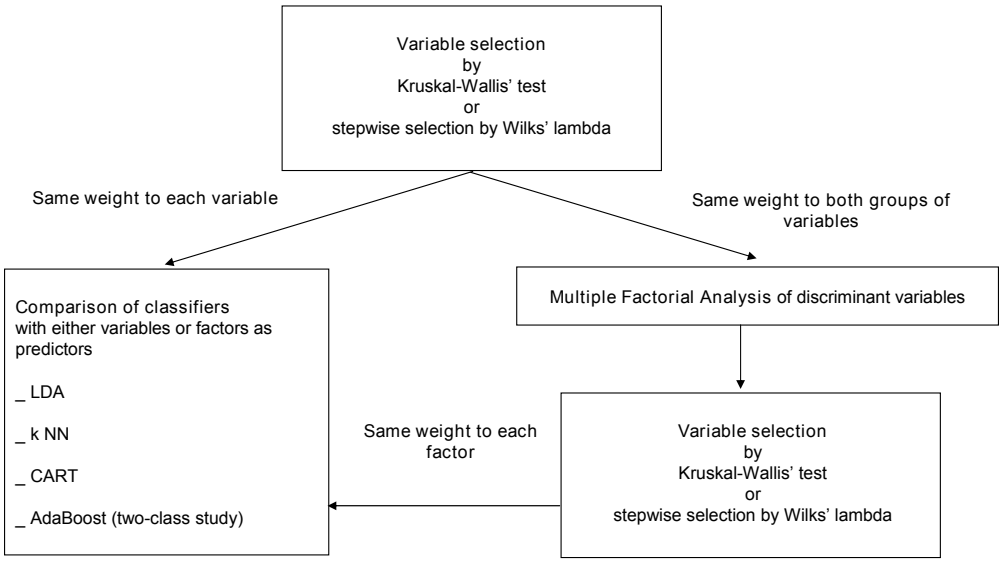

Figure 1: Specific statistical analysis for DBPCFC and first accident scores. The left path of the analysis gives a set of discriminant predictors without allowing for the weights of both groups of variables, whereas the right one uses equally-weighted groups of predictors.

Eliciting dose

To discriminate this variable, we also devised an algorithm that simultaneously clusters the eliciting dose values and selects predictors by minimizing the Wilks' lambda (Section 3.3). This algorithm was applied by using raw variables or factors computed by MFA with the 34 available variables as predictors. Once the predictors were chosen and the clusters built, the same statistical approaches as for DBPCFC and first accident scores were used. The corresponding statistical analysis is summarized in Figure 2.

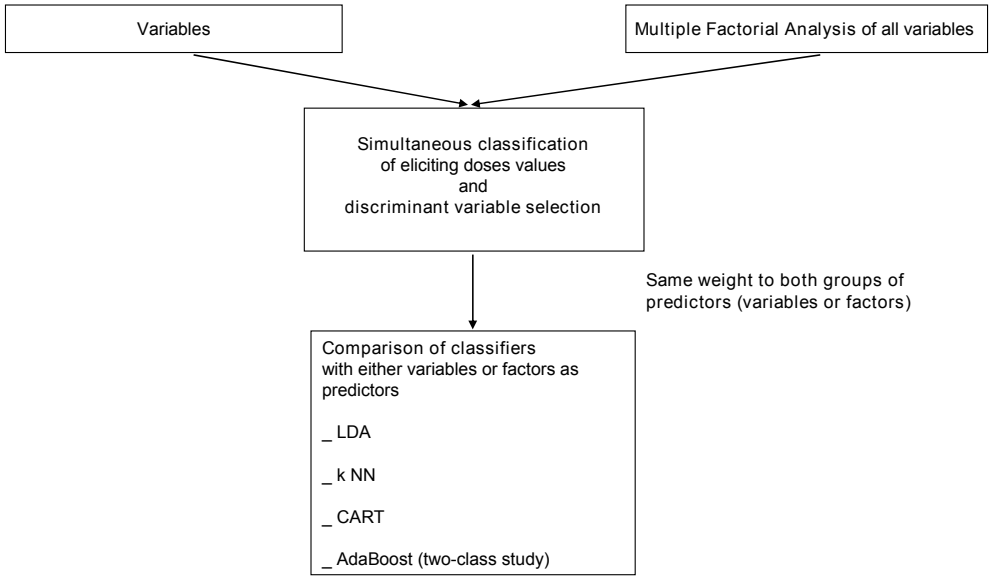

Figure 2: Specific statistical analysis for the eliciting dose

3.2 Multiple Factorial Analysis (MFA)

MFA was introduced by B.Escofier and J.Pagès [8, 9] for sensory analysis, and is a PCA with a particular choice of metric. The aim of this method is to give a similar part to several groups of variables when determining factors, *i.e.*, uncorrelated linear combinations of the initial variables. This procedure is useful for avoiding models that are fully influenced by a single group of numerous variables which could partially cancel the effect of the other groups. Briefly, an MFA is performed as follows :

Suppose p variables are measured on n subjects and divided in q groups :

$$(x^{1,1}, \ldots, x^{1,m_1}); \ldots; (x^{q,1}, \ldots, x^{q,m_q}) \tag{1}$$

with $\sum_{k=1}^{q} m_k$=p, where m_k is the number of variables in group k.

Denote $\mathbf{X}^k$ the matrix of data of size $n \times p$ corresponding to the k^{th} group of variables, namely :

$$\mathbf{X}^k = \begin{array}{c|ccccc} & x^{k,1} & \ldots & x^{k,j} & \ldots & x^{k,m_k} \\ \hline 1 & & & & & \\ \vdots & & & \vdots & & \\ i & & \ldots & x_i^{k,j} & \ldots & \\ \vdots & & & \vdots & & \\ n & & & & & \end{array}$$

where the generic element $x_i^{k,j}$ denotes the measure of the variable $x^{k,j}$ for the sample point i.

Let also $\mathbf{X} = (\mathbf{X}^1|\ldots|\mathbf{X}^q)$ be the matrix corresponding to the whole set of variables.

For the k^{th} group of variables, let $\mathbf{M}^k$ be a metric matrix in $\mathbb{R}^{m_k}$, k=1, ..., q.

Let $\mathbf{D}$ be the diagonal matrix of the weights assigned to the sample points.

The MFA algorithm is then as follows :

- Step 1 : For any $1 \leq k \leq q$, perform PCA($\mathbf{X}^k$,$\mathbf{M}^k$,$\mathbf{D}$), and denote λ_1^k the greatest eigenvalue corresponding to the first factor ;
- Step 2 : consider the metric matrix in $\mathbb{R}^p$:

$$\mathbf{M}= \begin{pmatrix} \mathbf{M}^1/\lambda_1^1 & & \\ & \ddots & \\ & & \mathbf{M}^{\mathrm{q}}/\lambda_1^{\mathrm{q}} \end{pmatrix}$$

  and perform PCA($\mathbf{X}$,$\mathbf{M}$,$\mathbf{D}$).

Note that in our case, $q = 2$ with the immunoassays as the first group of variables and the SPTs as the second. Variables were first centered and scaled to unity, and the metric matrix $\mathbf{M}^k$ was then set to the identity matrix $\mathbf{I}^k$ in $\mathbb{R}^{m_k}$.

For the DBPCFC and first accident scores, we thought it made more sense to compute the factors using only the discriminant variables rather than using all available variables. Indeed, the predictive model needed to be built with a reasonable number of characters. Even if a limited number of factors were chosen afterwards, all variables would still have to be measured to compute the factors. Moreover, a non-discriminant variable could have a large coefficient for some retained factors even though it would not improve the overall discriminative power of the model. Nonetheless, we did compute factors using all variables as well, but the results did not improve the discrimination of the first accident score. For the eliciting dose, factors were computed using all 34 available variables, not only the discriminant ones. As described in Section 3.3, the set of discriminant variables depends on the choice of the clustering of eliciting dose values and vice versa. Thus it did not seem appropriate to replace the optimal set of variables by factors.

3.3 An algorithm for simultaneously clustering the response variable and selecting discriminant variables

3.3.1 Principle

Eliciting dose is a variable whose values are taken according to an increasing scale of fixed doses of peanut. For a given patient, only an interval including the eliciting dose

is actually known. We wished first to group eliciting dose values in a limited number of intervals, and second to select the most discriminant variables for these categories. Here we propose an algorithm to perform these two steps simultaneously using alternate optimization.

For a given partition in intervals, a set of discriminant variables of fixed cardinal is selected using a certain optimality criterion. A new partition in intervals is then determined to optimize this criterion with the chosen variables, and so on. In order to not repeat this algorithm for different values of the cardinal, the number of variables to include could be increased one-by-one at each step of the procedure. The optimal set of variables could then be searched into all the possible subsets of variables of fixed cardinal corresponding to this step. To avoid heavy computations, variables were included forward in the model. At each step, a new variable was chosen according to the optimality criterion and added to the preceding variables.

Since Wilks' lambda provides a non-empirical stopping rule by testing its significance, this approach was preferred over using within-class inertia computation as the optimality criterion. The p-value of $F - to - enter$ was set to 0.15.

3.3.2 Constructing the clusters

Let y be an ordinal categorial variable of levels $\{m_1, \ldots, m_l\}$, $m_1 < m_2 < \ldots < m_l$. Suppose that we want to cluster the levels of y into a limited number r of intervals $]m_i, m_j]$. Only consecutive levels can be gathered. We build consecutive left-opened and right-closed intervals by selecting the upper bounds. Thus the number of possible clusterings is C_{l-1}^{r-1}.

3.3.3 Algorithm

The algorithm for computing intervals and selecting discriminant variables is as follows:

- Step 1 :
    1. choose the clustering $\mathcal{C}^1$ of eliciting dose values that minimises Λ, computed using all 34 available predictors ;
    2. select the predictor v^1 that minimises Λ with the obtained clustering $\mathcal{C}^1$;
- Step 2 :
    1. choose the clustering $\mathcal{C}^2$ of eliciting dose values that minimises Λ, computed using the previously selected predictor v^1 ;
    2. select the predictor v^2 such that the paired predictors (v^1, v^2) minimise Λ with the new clustering $\mathcal{C}^2$;

- and so on ...
- procedure stops if either no left predictor can improve the discriminant power of the model, *i.e.*, if $F-to-enter$ $p-$value is greater than 0.15 [13], or if every predictor is already entered.

Note that the $F-to-enter$ value is only computed when a new variable is entered into the model. This algorithm was used with both the variables and the MFA factors computed using all 34 available variables. Since new discriminant variables could have been chosen at each step of the algorithm, there was no default starting set of discriminant variables. Moreover, it did not seem appropriate to perform the MFA at the end of the algorithm, since the selected variables were specifically chosen to discriminate the found clusters.

3.4 Discriminant analysis

Linear discriminant analysis, k-NN, CART [10] and stepwise logistic regression [11] are classic methods. For two-class discrimination we also used the AdaBoost algorithm with CART [5] and penalized SVM. Since these are still recently developed algorithms, we briefly summarize their concepts below.

3.4.1 AdaBoost

Let $\{(\boldsymbol{x}_i, y_i)_{1\leq i\leq n}\}$ a dataset, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of predictors, and $y_i \in \{0,1\}$ is a binary response variable to discriminate. The principle of the AdaBoost algorithm is to re-weight observations that were misclassified by a base classifier (CART in our case). At each step of the procedure, a new classification tree is randomly built, inducing new misclassified sample points whose weights are updated before the following step starts. The method proceeds according to the following algorithm:

- Step 1 : assign equal weigths to all sample points $w_i^{[0]}=1/n, \forall i=1,\ldots,n$;
- Step 2 : for m=1, ..., M do :
  1. build a classifier $\hat{g}^{[m]}$ trained on data weighted by $w_i^{[m-1]}, \forall i=1,\ldots,n$;
  2. classify the data by resubstitution : determine $\hat{g}^{[m]}(\boldsymbol{x}_i), i=1,\ldots,n$;
  3. compute the misclassification rate :

$$err^{[m]} = \frac{\sum_{i=1}^{n} w_i^{[m-1]} \mathbb{1}_{(y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i))}}{\sum_{i=1}^{n} w_i^{[m-1]}}, \tag{2}$$

$$\alpha^{[m]} = \log\left(\frac{1-err^{[m]}}{err^{[m]}}\right), \tag{3}$$

where

$$\mathbb{1}_{(y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i))} = \begin{cases} 1 \text{ if } y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i) \text{ (\textit{i.e.} if } \boldsymbol{x}_i \text{ is misclassified)}, \\ 0 \text{ otherwise;} \end{cases} \tag{4}$$

4. update the weights

$$\tilde{w}_i = w_i^{[m-1]} \exp(\alpha^{[m]} \mathbb{1}_{(y_i \neq \hat{g}^{[m]}(\boldsymbol{x}_i))}), \tag{5}$$

$$w_i^{[m]} = \frac{\tilde{w}_i}{\sum_{j=1}^{n} \tilde{w}_j}; \tag{6}$$

- Step 3 : build the aggregated classifier

$$\hat{f}_{\text{AdaBoost}}(\boldsymbol{x}) = \arg\max_{y \in \{0,1\}} \sum_{m=1}^{M} \alpha^{[m]} \mathbb{1}_{(\hat{g}^{[m]}(\boldsymbol{x})=y)}. \tag{7}$$

A novel observation $\boldsymbol{x}$ is classified by the majority vote $\hat{f}_{\text{AdaBoost}}(\boldsymbol{x})$, where vote m is weighted by $\alpha^{[m]}$.
In this study AdaBoost was performed for $M = 50, 100$ and 200 on the training set and the parameter value giving the best result was kept.

3.4.2 Penalized SVM

Let $\{(\boldsymbol{x}_i, y_i)_{1 \leq i \leq n}\}$ a training set, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is the vector of predictors, and $y_i \in \{-1, 1\}$ is the class label. The Support Vector Machine (SVM) algorithm gives the hyperplane H that best splits both groups and that is defined by the equation

$$H : f(\boldsymbol{x}) = < \boldsymbol{w}, \boldsymbol{x} > + b = 0 \tag{8}$$

where $< ., . >$ is the usual dot product in $\mathbb{R}^p$, $\boldsymbol{w}$=(w^1,...,w^p) are the coefficients of the hyperplane and b is the intercept.
The coefficients are obtained by solving the convex optimization problem :

$$\min_{b, \boldsymbol{w}} \sum_{i=1}^{n} [1 - y_i f(\boldsymbol{x}_i)]_+ + \text{pen}_\lambda(\boldsymbol{w}) \tag{9}$$

where λ is a positive tuning parameter, $[.]_+ = \max(., 0)$ is the positive part and $\text{pen}_\lambda(\boldsymbol{w}) = \lambda \|w\|_2^2 = \lambda \sum_{j=1}^{p} (w^j)^2$.

The class label of a novel observation $\boldsymbol{x}$ is then given by $\text{sign}(f(\boldsymbol{x}))$.

To select a limited number of variables, the term pen_λ in (9) can be replaced by a penalization function being singular at the origin and having a continuous first-order derivative [22]. Two functions were used in this study :

- L_1 : $\text{pen}_\lambda(\boldsymbol{w}) = \lambda\|\boldsymbol{w}\|_1 = \lambda\sum_{j=1}^{p}|w^j|$
- Smoothly Clipped Absolute Deviation (SCAD) : $\text{pen}_\lambda(\boldsymbol{w}) = \sum_{j=1}^{p} p_\lambda(w^j)$ where

$$p_\lambda(w^j) = \begin{cases} \lambda|w^j| & \text{if } |w^j| \leq \lambda \\ -\frac{(|w^j|^2 - 2a\lambda|w^j| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |w^j| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |w^j| > a\lambda \end{cases} \tag{10}$$

  where $a > 2$ is a tuning parameter. As suggested in [2], a is set by default to 3.7.

The optimal λ was chosen by the algorithm in the set $\{0.05, 0.10, 0.15, \ldots, 0.95\}$ for L_1 penalization and in $\{0.10, 0.20, \ldots, 1\}$ for SCAD penalization.

For both penalizations, variables with coefficient $|w^j|$ lower than a given ϵ were considered useless and removed from the model. In the penalized SVM R package, ϵ is set to 0.001 [22].

4 Results

4.1 Principal Component Analysis

The PCA representation of the variables was relevant. As seen on the correlation circle of Figure 3, intra-family correlations between variables were rather high but inter-family correlations were quite low (with the notable exception of nuts and aeroallergens). Moreover, the *total IgE* and *specific IgE to rAra-h8* did not seem closely related to the other immunoassays, an observation fully validated by clinicians. Indeed, as explained earlier, the level of *total IgE* is the global measure of this antibody subclass, whereas the *specific IgE to rAra-h1, rAra-h2* and *rAra-h3* are directed against peanut allergens alone. Also, *rAra-h8* is a homologous protein to the birch pollen allergen *Bet-v1*, sharing about 66% of their amino acid sequences. Thus patients sensitive to both peanut and birch pollen could present high values of *specific IgE to rAra-h8* without being allergic to peanuts.

These results confirmed clinical observations. Also, the individual representation did not provide supplementary information (data not shown), and no particular interpretation was evident for the PCA axes.

4.2 Discriminant analysis

Here we show the results for the prediction of the first accidental exposure score, the DBPCFC score and the eliciting dose. Tables 3, 4 and 5 give the well-classification rates obtained by combining the classifiers with the 3 different statistical approaches :

1. direct application of the classifiers,

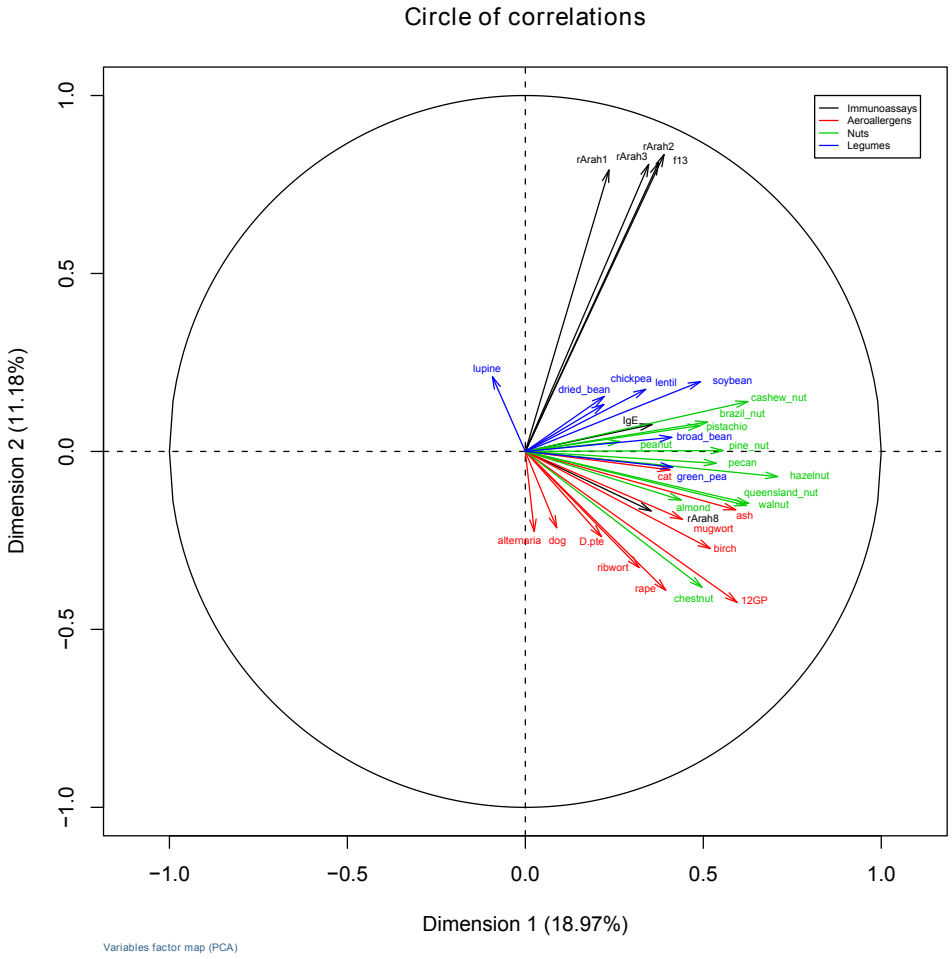

Figure 3: Circle of correlations

2. simultaneous variable selection and classification methods,

3. specific variable selection scheme prior to classification.

The percentage of detected patients with high severity was also computed. According to clinicians, failing to detect patients with severe allergies could indeed lead to inappropriate food intake by the patient.

Results are given including the variable selection scheme and/or MFA in the cross-validation. For the first and the third variable selection schemes, only the classifier giving the best performances among LDA, $k-$NN, CART and AdaBoost is displayed. For penalized SVM, the best result between L_1 ans SCAD penalizations was kept.

The variables obtained with the selection process offering the bests results are also given and summarized in Figure 4 for each measure of severity .

4.2.1 First accidental exposure score

Four-class study

Recall that in what follows, we combined scores 4 and 5 because of the low frequency of score 5. Thus the four classes considered here are for scores of $\{1\}$, $\{2\}$, $\{3\}$, and $\{4, 5\}$.

Well-classification rates obtained by fourth-fold cross-validation with each statistical approach are shown in Table 3. The percentage of patients of score 4 who were correctly classified in class 4 is also given. Note that direct application of the classifiers could not have been performed since the number of variables was greater than the size of the learning sets. $5-$NN combined with our specific variable selection scheme was the most performant classifier with 41% of well-classified patients, since stepwise logistic regression and penalized SVM did not give better results. Nevertheless all these results remained poor. Replacing variables by factors did not give better results than direct use of the variables.

On the whole dataset 7 variables had a Kruskal-Wallis p-value lower than 0.10 (*peanut, walnut, chick pea, pecan nut, broad bean, green pea, cashew nut*, ordered by increasing p-value, Table 6) and 8 were retained in the Wilks' lambda selection (*chick pea, specific IgE to rAra-h8, green pea, rape seed, peanut, ribwort, total IgE, specific IgE to rAra-h1*, Table 7). Thus 12 different variables were selected in total (3 immunoassays and 9 SPTs).

Two-class study

The methodology used was the same as for the four-class study.

The method which gave the best results for two-class discrimination was $1-$NN with factors computed from discriminant variables as predictors (81% of well-classified sample points) (Table 3). Of the score 4 patients, 74% were correctly classified. Contrary to the four-class study, the use of factors improved the classification rates compared to the direct use of variables since using selected variables with $1-$NN gave 79% of well-classification rate. The other models yielded poor results.

Processing the variable selection on the whole dataset gave interesting results. Eleven different variables were selected to build discriminant factors: *peanut, walnut, specific IgE to rAra-h1, specific IgE to rAra-h3, pecan nut* for Kruskal-Wallis and *peanut, specific IgE to rAra-h1, lupine flour, specific IgE to rAra-h2, ribwort, Dpte, birch, dog epithelia* for Wilks' lambda. These immunoassays, as well as SPTs to nuts and legumes, are variables expected by clinicians. This indicates that our selection process seems to detect "useful" variables. More surprisingly, a few SPTs to aeroallergens were also selected. Indeed, these variables are not known to cross-react with peanuts. Note that the discriminant factors will have to be computed these variables to use the best model for further classification.

4.2.2 DBPCFC score

The statistical approach used to predict the DBPCFC score was the same as for the first accident score.

Four-class study

Scores of 4 and 5 were again grouped together due to the low frequency of score 5.

Although applying CART without preselecting variables gave the best results with a 38% successful classification rate, the overall misclassification error remained high (Table 4). Note that with the specific variable selection approach, MFA was not performed since only SPTs were entered in the model during cross-validation.

Interestingly 7 variables were entered in the classification tree performed on the whole dataset : *lupine flour, specific IgE to f13, lentil, total IgE, specific IgE to rAra-h3, 12 grass pollens, specific IgE to rAra-h8*, in order of selection. This means that although all variables were available for building a model only a few were considered as useful by the CART algorithm.

Two-class study

DBPCFC scores can be gathered into two classes in the same manner as for the first accidental exposure score. Overall, using 3-NN classification with our variable selection scheme offered the best results, with 66% of successfully classified sample points and 33% of successfully classified severe patients (Table 4).

Total IgE was the only immunoassay retained in the model during cross-validation. Thus, no factor was computed. Indeed, assigning the same weight to this single variable as to the other variables did not seem appropriate, because *total IgE* are not antibodies specifically involved in peanut allergies, but in all allergic reactions.

On the whole dataset this resulted in selecting variables *lupine flour* for Kruskal-Wallis and *almond, total IgE, lupine, broad bean, pine nut* for Wilks' lambda.

4.2.3 Eliciting dose

Before applying directly all the classifiers without selecting variables and performing stepwise logistic regression and penalized SVM, eliciting dose was first converted into a class variable. During cross-validation values were gathered into the partition of minimal Wilks' lambda computed with all available variables. These results were compared to those obtained with our algorithm that simultaneously selects discriminant variables and groups the eliciting dose values in an optimal partition (Section 3.3). This was performed for both four-class ($r = 4$) and two-class ($r = 2$) studies. The minimal number of sample points in each class was set to 10, in order to have class frequencies large enough to perform cross-validation.

Note that this process was included in cross-validation.

Four-class study

The eliciting dose was correctly predicted for 39% of the patients using our algorithm and CART with variables. Moreover, only 52% of the patients from the lowest eliciting doses group were correctly classified. Other approaches did not give better results (Table 5) .

Table 8 shows the bounds x, y, z for clustering $]-\infty, x],]x, y],]y, z],]z, +\infty[$, and the variables entered at each step of the algorithm when performed on the whole dataset. The algorithm stopped at step 7 because no additional improvement resulted afterwards. The selected variables were *hazelnut, birch, pistachio, cashew nut, green pea, total IgE, pine nut* and the bounds were 95, 215, and 500 *mg*.

The same study was performed with factors as predictors but it did not enhance the model.

Two-class study

Selecting factors of all available variables with our algorithm was the most discriminant model to discriminate the eliciting dose in the two-class study (77% were successfully classified with $5-$NN). It also successfully classified 72% of the highly reactive patients. The threshold of eliciting dose was 300 *mg* with 7 factors selected ($\Lambda = 0.62$). Nevertheless, this model cannot easily be used in practice. Indeed, as mentioned earlier, MFA was performed on all 34 variables, not just the discriminant variables. This means that to correctly predict the eliciting dose, all variables would have to be measured to compute the factors; this does not seem feasible.

5 Conclusions

To the best of our knowledge, this paper presents the first discriminant analysis of DBPCFC score, first accident score, and eliciting dose measurement of peanut allergy severity. Previous studies were aimed at finding links between immunoassays or SPTs and allergy severity, but their statistical analyses were limited to either comparing distributions between groups of patients (using, for instance, the Mann-Whitney test) or to computing linear correlation coefficients [12, 19]. In our approach, we used several classification rules to aid in comparing and choosing the optimal and most efficient method. It appeared that in general selecting discriminant variables by a process independent from classification gave better results. In addition, we found that using MFA to compute new predictors was an attractive solution when equalizing the weights of group variables, and we proposed a novel algorithm for simultaneously clustering the levels of ordinal qualitative variables and for selecting discriminant variables.

Our work differs from earlier studies in several respects. Previous studies were performed on small sample sizes of 30 to 40 patients [12, 19] using a small number of measured variables, which only permitted a limited choice of discriminating predictors. Also, *specific IgE to Ara-h1,2,3* were measured by SPTs instead of immunoassays [19] and although a positive response to an SPT does indeed indicate allergen sensitivity, it is still less accurate than immunoassays.

Figure 4: Summary of discriminant variables. Selected variables are colored in black for each measure of severity. In two classes, all available variables were used to compute discriminant factors of eliciting dose

There are several scoring methods in the literature to evaluate peanut allergy severity. For example, Hourihane et al. devised a complex 25-class scoring system combining observed reactions and eliciting doses [12]. A graduation of symptoms was also proposed by Müller [18], but this score is based on allergic reactions in response to bee or wasp venom (and not peanut allergens). Thus, developing a standardized scoring method is still necessary and would facilitate comparison studies from different centers. One possible solution would be comparing the results of Hourihane et al.'s score with the one used in this study using the same cohort of patients. Moreover, the large number of scoring levels in Hourihane et al.'s approach could be reduced by our algorithm.

Nevertheless, several potential biases in our study must be noted. First, SPT diameters are relatively imprecise. There is no standard method of measuring SPT reaction since both wheal diameters and areas are popular metrics [19]. Additionally, the score of the first accident can be inaccurate or imprecise since it depends on the medical history of the patient, and hence subject to inaccuracies in the patient's memory which may underestimate symptom severity. Other factors such as medication can affect the first reaction symptoms as well [15, 17] yielding a severity score higher than it should be. Finally, the first exposure score is a past event predicted using variables measured during

a subsequent DBPCFC. As mentioned in Section 4.2.2, Hourihane et al. also used such a reverse prediction with a community score that was evaluated *a posteriori* using the record file of the patient [12].

The predictive models used in our study yielded correct classifications for the first accidental exposure and DBPCFC of up to 81% and 66% for the two-class study. Our algorithm also allowed us to group eliciting dose values and to select discriminant predictors, leading to an 77% classification rate for the two-class study. This indicates that it is indeed possible to correctly predict peanut allergy severity by measuring well-chosen variables. Considering that all immunoassays of specific *IgE* were selected once, we also hypothesize that measuring new antibodies to peanut allergens, such as those directed against *rArah-6, rArah-7* and *rArah-9*, will further improve the discriminative power of our models. This also argues for these antibodies playing key roles in the diagnosis of peanut allergy severity.

Our variable selection process also offers a new perspective on conducting allergy check-ups. Indeed, some unexpected variables appeared several times in our models, such as SPTs to *dog epithelia* as shown in Figure 4. If future experiments could distinguish the medical relevance of this observation from cross-reactivity, then the importance of these SPTs in discriminating severity would be confirmed. Besides, some SPTs never appeared in our models, such as SPTs to *Alternaria* or *Brazil nut*, and thus should no longer be performed in practice when diagnosing peanut allergy severity. Furthermore, some SPTs with proven cross-reactivity to peanuts were retained in our models, such as *lupine flour* [16], indicating that our results were in line with other medical discoveries.

The discriminating models described in this paper are a first step towards a simple, safe and efficient diagnosis of peanut allergy severity by quantifying antibodies. Before being applied in clinical practices, they must first be validated on an independent set of patients. New variables must also be added as additional predictors toward improving successful classification rates. These models could then become practical tools for clinicians. When scoring severity, the clinical test results could be reported online at the allergy vigilance network (*Réseau d'allergovigilance, http://www.cicbaa.com/*), or a simple statistical software could be programmed.

Acknowledgements

The authors would like to thank Frances T. Yen for her critical review of the manuscript.

References

[1] C. Astier, M. Morisset, O. Roitel, F. Codreanu, S. Jacquenet, P. Franck, V. Ogier, N. Petit, B. Proust, D.A. Moneret-Vautrin, A.W. Wesley Burks, B.E. Bihain, H.A.

Sampson, and G. Kanny, *Predictive value of skin prick tests using recombinant allergens for diagnosis of peanut allergy*, Journal of Allergy Clinical Immunology 118 1 (2006), pp. 250–256.

[2] N. Becker, W. Werft, G. Toedt, P. Lichter and A. Benner, *penalizedSVM: a R-package for feature selection SVM classification*, Bioinformatics 25 (2009), pp. 1712–1712.

[3] C. Bindslev-Jensen, B. Ballmer-Weber, U. Bengtsson, C. Blanco, C. Ebner, J. Hourihane, A.C. Knulst, D. Moneret-Vautrin, K. Nekam, B. Niggemann, M. Osterballe, C. Ortolani, J. Ring, C. Schnopp and T. Werfel, *Standization of food challenges in patients with immediate reactions to foods-position paper from the European Academy of Allergology Clinical Immunology*, Allergy 59 (2004), pp. 690–697.

[4] S.A. Bock, A. Munoz-Furlong, H.A. Sampson, *Fatalities due to anaphylactic reactions to foods*, Journal of Allergy Clinical Immunology 107 (2001), pp. 191–193.

[5] P. Buhlmann and T. Hothorn, *Boosting Algorithms: Regularization, Prediction Model Fitting*, Statistical Science 22 (2007), pp. 477–505.

[6] W.J. Conover, *Practical Nonparametric Statistics*, John Wiley & Sons, New-York, 1980.

[7] M.C Costanza and A.A. Afifi, *Comparison of stopping rules in forward stepwise discriminant analysis*, Journal of the American Statistical Association 74 (1979), pp. 777–785.

[8] B. Escofier and J. Pagès, *Multiple factor analysis (AFMULT package)*, Computational Statistics Data Analysis 18 (1994), pp. 121–140.

[9] B. Escofier and J. Pages, *Analyses factorielles simples et multiples, 3ème édition* Dunod, Paris, 1998.

[10] T. Hastie, R. Tibshirani, and J. Friedman *The elements of statistical learning: data mining, inference and prediction*, Springer-Verlag, New-York, 2001.

[11] D.W. Hosmer and S. Lemeshow *Applied logistic regression*, John Wiley & Sons, New-York, 2000.

[12] J.O.B. Hourihane, K.E.C. Grimshaw, S.A. Lewis, R.A. Briggs, J.B. Trewin, R.M. King, S.A. Kilburn and J.O. Warner, *Does severity of low-dose, double-blind, placebo-controlled food challenges reflect severity of allergic reactions to peanut in the community?*, Clinical & Experimental 35 (2005), pp. 1227–1233.

[13] R.I. Jennrich, *Stepwise Discriminant Analysis*, in Statistical Methods for Digital Computers, K. Enslein, A. Ralston, and H. Wilf, eds., John Wiley & Sons, New York, 1977 , pp. 76–96.

[14] P.A. Lachenbruch, *Discriminant analysis*, Hafner Press, Collier Macmillan Publishers, London, 1975.

[15] J. Lidholm, B.K. Ballmer-Weber, A. Mari and S. Vieths, *Component-resolved diagnostics in food allergy*, Current Opinion in Allergy Clinical Immunology 6 (2004), pp. 234–240.

[16] D.A. Moneret-Vautrin, L. Guerin, G. Kanny, J. Flabbee, S. Frémont and M. Morisset, *Cross-allergenicity of peanut lupin : the risk of lupin allergy in patients allergic to peanuts*, Journal of Allergy Clinical Immunology 104 (1999), pp. 883-888.

[17] D.A. Moneret-Vautrin, M. Morisset, J. Flabbee, E. Beaudouin and G. Kanny, *Epidemiology of life-threatening lethal anaphylaxis: a review*, Allergy 60 (2005) pp. 443–451.

[18] H.L. Müller, *Diagnosis Treatment of Insect Sensitivity*, Journal of Asthma 3 (1966), pp. 331–333.

[19] K. Peeters, S.J. Koppelman, E. van Hoffen, C.W.H. van der Tas, C.F. den Hartog Jager, A.H. Penninks, S.L. Hefle, C. Bruijnzeel-Koomen, E.F Knol and A.C. Knulst, *Does skin prick test reactivity to purified allergens correlate with clinical severity of peanut allergy?*, Clinical & Experimental Allergy 37 (2007), pp. 108–115.

[20] T.D. Pollard, W.C. Earnshaw and G.T. Johnson, *Cell biology*, Saunders Philadelphia, 2004.

[21] R Development Core Team, *R: A Language Environment for Statistical Computing*, 2007, software available at *http://www.R-project.org*.

[22] H.H. Zhang, J. Ahn, X. Lin and C. Park, *Gene selection using support vector machines with non-convex penalty*, Bioinformatics 22 (2006), pp. 88-95.

Table 1: Frequencies of DBPCFC and first accidental exposure severity score

scores	1	2	3	4	5	sum
DBPCFC	17	25	11	22	1	76
first accident	9	16	4	16	2	47

Table 2: Frequencies of eliciting dose values

eliciting dose (mg)	n	eliciting dose (mg)	n
1.4	1	300	1
4.4	2	400	1
14	1	500	18
15	2	965	4
44	8	1000	1
65	5	2000	3
95	1	2110	1
115	1	3500	1
165	1	3610	1
210	1	4110	1
215	17	7000	3
265	1	sum	76

Table 3: Results of the discriminant analysis of first accidental exposure score for the three kinds of variable selection schemes. The best classification method is given for the four-class and two-class studies. Results are expressed as successful classification rates and as severe patient detection rates.

method	4 classes	2 classes
no selection	x	AdaBoost ($M = 50$) : 63%-33%
stepwise logistic regression	34%-65%	63%-22.5%
penalized SVM	x	L_1 : 61%-44%
Kruskal-Wallis / Wilks' lambda with variables	5NN : 41%-65%	1NN : 79%-72%
Kruskal-Wallis / Wilks' lambda with factors	1NN : 34%-46%	1NN : 81%-74%

Table 4: Results of the discriminant analysis of DBPCFC score for the three kinds of variable selection schemes. The best classification method is given for the four-class and two-class studies. Results are expressed as successful classification rates and as severe patient detection rates.

method	4 classes	2 classes
no selection	CART : 38%-44%	2NN : 64%-37%
stepwise logistic regression	35%-32%	56%-10%
penalized SVM	x	L_1 : 62%-24%
Kruskal-Wallis / Wilks' lambda with variables	2NN : 36%-52%	3NN : 66%-33%
Kruskal-Wallis / Wilks' lambda with factors	x	x

Table 5: Results of the discriminant analysis of eliciting dose for the three kinds of variable selection schemes. The best classification method is given for the four-class and two-class studies. Results are expressed as successful classification rates and as severe patient detection rates.

method	4 classes	2 classes
no selection	LDA : 38%-39%	5NN : 69%-62%
stepwise logistic regression	12%-5%	58%-48%
penalized SVM	x	L_1 : 66%-45%
algorithm with variables	CART : 39%-52%	1NN : 66%-36%
algorithm with factors	LDA : 33%-34%	5NN : 77%-72%

Table 6: p-values of Kruskal-Wallis tests for first accidental exposure score (four-class study). Variables are ordered by increasing p-values.

variable	p-value	variable	p-value
peanut	0.023	total IgE	0.424
walnut	0.038	chest nut	0.436
chick pea	0.051	ribwort	0.442
pecan nut	0.060	12 grass pollens	0.455
broad bean	0.067	lentil	0.474
green pea	0.081	mugwort	0.488
cashew nut	0.088	sp.IgE to f13	0.498
sp.IgE to rAra-h3	0.146	sp.IgE to rAra-h8	0.570
sp.IgE to rAra-h2	0.160	rape seed	0.582
soybean	0.172	ash	0.631
pistachio	0.201	sp.IgE to rAra-h2	0.693
Alternaria	0.216	pine nut	0.698
dried bean	0.254	Dpte	0.706
Brazil nut	0.266	birch	0.723
dog ephitelia	0.291	almond	0.736
Queensland nut	0.317	hazelnut	0.766
lupine flour	0.372	cat ephitelia	0.806

Table 7: Wilks' lambda stepwise selection for first accidental exposure score (four-class study).

step	entered	Wilks' Λ	$F-to-enter$	$F-to-enter$ p-value
1	chick pea	0.74	5.14	4.00E-03
2	Sp.IgE to rAra-h8	0.55	4.84	5.55E-03
3	green pea	0.39	5.36	3.31E-03
4	rape seed	0.31	3.76	1.81E-02
5	peanut	0.26	2.50	7.41E-02
6	ribwort	0.21	2.56	6.95E-02
7	Total IgE	0.18	2.09	1.18E-01
8	Sp.IgE to rAra-h1	0.15	2.24	9.98E-02

Table 8: Clusters and variables selected to discriminate eliciting doses and corresponding Λ and $F-to-enter$ statistics (four-class study).

step	bounds / predictors	Wilks Λ	$F-to-enter$	$F-to-enter$ p-value
1	95-215-500	0.08		
1	hazelnut	0.73	8.51	6.96E-05
2	95-215-500	0.73		
2	birch	0.64	3.15	0.031
3	95-215-1000	0.63		
3	pistachio	0.56	2.85	0.044
4	95-215-1000	0.56		
4	cashew nut	0.50	2.74	0.050
5	95-215-1000	0.50		
5	green pea	0.45	2.24	0.092
6	95-215-1000	0.45		
6	total IgE	0.41	2.20	0.096
7	95-215-500	0.41		
7	pine nut	0.35	3.24	0.028