Research Article

Variable Selection Method in Prediction Models: Application in Periodontology

Paul Tramini, 1 Jean-Christophe Chazel, 2 Isabelle Calas-Bennasar, 3 Philippe Gibert, 3 and Nicolas Molinari 4

1 Dental Public Health, Faculty of Odontology, University of Montpellier 1, 545 avenue du Pr. JL Viala, 34193 Montpellier Cedex 5, France
2 Restorative Dentistry, Faculty of Odontology, University of Montpellier 1, 545 avenue du Pr. JL Viala, 34193 Montpellier Cedex 5, France
3 Periodontology, Faculty of Odontology, University of Montpellier 1, 545 avenue du Pr. JL Viala, 34193 Montpellier Cedex 5, France
4 Biostatistics, Faculty of Medicine, University of Montpellier 1, 545 avenue du Pr. JL Viala, 34193 Montpellier Cedex 5, France

Correspondence should be addressed to Paul Tramini; paul.tramini@orange.fr

Received 27 September 2013; Revised 16 December 2013; Accepted 23 December 2013; Published 4 February 2014

Academic Editor: Hideki Ohyama

Copyright © 2014 Paul Tramini et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The aim of this study, applied in the field of periodontal diseases, was first to analyze the fatty acid levels in two groups of patients and then to propose a method for selecting the most relevant predictors. Two groups of patients, 29 with moderate or severe periodontitis and 27 who served as controls, were clinically examined, and their fatty acids in serum were measured by gas chromatography. The levels of these 12 fatty acids were the variables of the analysis. Logistic regression, together with the area under the receiver operating characteristic (ROC) curves, allowed determining a composite score which led to a subset of the most relevant covariables. The fatty acid levels differed significantly between the 2 groups in multivariate analysis ($P = 0.03$) and the best logistic model was obtained with only 3 predictive variables: arachidonic acid, linoleic acid, and DHA. Fatty acid levels in serum of patients were significantly different according to the presence of moderate or severe periodontitis. By taking into account the comparison of ROC curves, our approach could optimize the choice of variables in multivariate analyses and could better fit it with diagnosis and prognosis of oral diseases in dental research.

1. Introduction

Periodontal disease is one of the most common causes of tooth loss among adults [1, 2]. It is characterized by chronic inflammation, periodontal pocket formation together with alveolar bone, and gum destruction. It is now admitted that bacteria alone are insufficient to cause such diseases and that the host’s characteristics are indeed known to be determining as well: heredity, tobacco, systemic diseases, and nutrition [3, 4]. So the aetiological role of systemic factors is difficult to separate from that of local factors [5].

It has been demonstrated that the balance between polyunsaturated and saturated serum fatty acids (FA) played an important role in the bone remodelling of the skeleton [6, 7]. So alveolar bone and periodontal tissues can be affected in the same way. Because some polyunsaturated FA are the precursors of prostaglandins which are mediators of inflammation, higher levels are associated with inflamed tissues [8, 9]. Thus, fatty acid measurements could be correlated to periodontal disease, such as periodontitis and alveolar bone loss. Moreover, in this study, we investigated the FA concentrations in the serum of two groups of patients: with periodontitis and without (or with mild) periodontitis. The aim of this study, applied in the field of periodontal diseases, was first to analyze the FA levels in two groups of patients and then to propose a method for selecting the most relevant predictors among a set of quantitative variables in a prediction model. This new approach was developed in order to simplify and optimize the choice of predictors, among the numerous potential predictive symptoms and biological analyses already described in periodontal research.
2. Materials and Methods

2.1. Patients’ Recruitment. The subjects were recruited in the dental hospital of Montpellier (France), from September 2010 to June 2011. The inclusion criteria comprised age from 35 to 55 years old and at least the presence of 20 natural teeth per subject. The patients who agreed with this study were interviewed and their general health was ascertained. Exclusion criteria were smoking, chronic systemic disorders, or prescription of systemic antibiotics or anti-inflammatory or other systemic drugs, because they may influence the relationship between the FA levels and the dependent variable.

2.2. Clinical Examinations. Clinical examinations were performed by one calibrated dentist (P.G.) in the Department of Periodontology of the Dental Hospital of Montpellier. Probing pocket depth (PPD) and clinical attachment loss (CAL) were assessed at six sites per tooth with a probe PCPUNC157 (Hu-Friedy, Chicago, USA). PPD was measured as the distance from the gingival margin to the base of the gingival sulcus or periodontal pocket, and CAL was the distance from the cementoenamel junction to the base of the sulcus or periodontal pocket. The periodontal status was determined as proposed by Page and Eke [10]:

(i) severe periodontitis: two or more interproximal sites with CAL ≥ 6 mm, not on the same tooth, and one or more interproximal sites with PPD ≥ 5 mm;

(ii) moderate periodontitis: two or more interproximal sites with CAL ≥ 4 mm, not on the same tooth, or two or more interproximal sites with PPD ≥ 5 mm;

(iii) no or mild periodontitis: neither moderate nor severe periodontitis.

After clinical examinations, two groups of patients were determined. The first group (group 1), was composed of 27 subjects who were diagnosed with mild or no periodontitis. The mean age was 43.4 ± 4.6 years (5 men and 22 women). There were 4 men and 18 women without periodontitis and one man and 4 women with mild periodontitis. The percentages of affected sites per subject were, respectively, 6.8% (2-3 mm), 0% (4-6 mm), and 0% (≥7 mm). The second group (group 2) comprised 29 subjects (6 men and 23 women) suffering from moderate or severe periodontitis, mean age: 44.5 ± 5.2 years. Moderate periodontitis was present in 4 men and 10 women, while severe periodontitis was present in 2 men and 13 women. The distribution of PPD was 36.2% (2-3 mm), 21.8% (4-6 mm), and 6.2% (≥7 mm). A randomly chosen sample of five subjects in each group (17% of the whole sample) was reexamined by the same dentist. The intraindividual reproducibility of CAL and PPD measurements was very good, with a Cohen kappa equal to 0.87 and 0.90, respectively. After written consent from all the subjects (groups 1 and 2) to use their blood samples and undergo clinical oral examinations, their FA were measured in serum by gas chromatography and expressed in grams per litre. This noninterventional study did not require approval by an ethics committee, since the blood samples (5 mL) were drawn from a database of patients consulting check-ups before oral surgery under general anaesthesia, or for regular check-ups in the department of preventive medicine. Blood samples were drawn after an overnight fast from the antecubital vein in nonheparinised test tubes and centrifuged at 1500 g for 30 min at 48°C. Triglycerides, cholesterol esters, and phospholipids were isolated by thin-layer chromatography. Hydrogen was the carrier gas. Results are given as percentage of total moles of FA. Seven unsaturated fatty acids and five saturated acids were measured in this study. So a total of 12 variables were used in the statistical analyses.

2.3. Univariate Analysis. Descriptive data were summarized as mean ± standard deviation (SD) and coefficient of variation (CV). According to the normality of the distribution, assessed with the Shapiro-Wilk test, comparison of fatty acid levels between the two groups was tested by univariate analyses (Student’s t-test or Mann-Whitney test). Relative differences between group 1 and group 2, expressed as a percentage of the mean value among diseased patients, were also calculated in order to summarize the variation of FA concentration.

2.4. Multivariate Analysis. The relationship between the FA levels and the response variable (with/without periodontitis) was fitted using logistic regression after calibration by the Hosmer-Lemeshow goodness-of-fit test. Multicollinearity was tested by estimating the variance inflation factor (VIF). Every variable associated with a P value below 0.20 in the univariate analysis was entered in the logistic model. A forward procedure was used to select the final multivariate model. FA effects are expressed as odds ratio with 95% confidence interval. Logistic regression is the most commonly used method to assess the relationship between one or more independent variables (the FA levels) and a binary response variable (with/without periodontitis).

The obtained logistic model allowed proposing a quantitative composite score. This score summarizes the effects of the FA levels and can be defined as a new predictor. Because the new predictor is linear combination of the fatty acid levels, it is a continuous variable. A standard approach to summarize the predictor performance was to examine all possible cutpoints. Each cutpoint yielded an estimated sensitivity and specificity. Sensitivity and specificity are, respectively, the chance that a true positive and a true negative will be identified as such by the predictor. A good predictor ought to have high values for both sensitivity and specificity [11]. The receiver operating characteristic (ROC) curve was then defined. The area under the ROC curve (AUC) was used as a global summary statistic of predictive accuracy. The optimal threshold cutpoint was determined from the ROC curve.

A cross-validation technique was used to assess how the results will generalize to an independent data set. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset, called the training set, and validating the analysis on the other subset, called the validation set or testing set. To reduce variability, multiple rounds of cross-validation were performed using different partitions, and the validation
Table 1: Simple statistics for the fatty acid levels in the two groups.

<table>
<thead>
<tr>
<th></th>
<th>Group 1 (n = 27)</th>
<th>Group 2 (n = 29)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Myristic</td>
<td>1.153</td>
<td>0.402</td>
</tr>
<tr>
<td>Palmitic</td>
<td>23.402</td>
<td>3.593</td>
</tr>
<tr>
<td>Palmitoleic</td>
<td>2.507</td>
<td>1.159</td>
</tr>
<tr>
<td>Searic</td>
<td>6.955</td>
<td>1.113</td>
</tr>
<tr>
<td>Oleic</td>
<td>18.519</td>
<td>2.615</td>
</tr>
<tr>
<td>Linoleic</td>
<td>30.055</td>
<td>3.421</td>
</tr>
<tr>
<td>Gamma-Linolenic</td>
<td>0.386</td>
<td>0.157</td>
</tr>
<tr>
<td>DHGL</td>
<td>1.279</td>
<td>0.373</td>
</tr>
<tr>
<td>Arachidonic</td>
<td>5.711</td>
<td>1.153</td>
</tr>
<tr>
<td>Alphalinolenic</td>
<td>0.406</td>
<td>0.220</td>
</tr>
<tr>
<td>EPA</td>
<td>0.748</td>
<td>0.341</td>
</tr>
<tr>
<td>DHA</td>
<td>1.852</td>
<td>0.718</td>
</tr>
</tbody>
</table>

Unit: percentage of total moles of FA; DHGL: dihomogammalinolenic acid; EPA: eicosapentaenoic acid; DHA: docosahexaenoic acid.

(* Significant outcome for Student’s t-test.)

Table 2: Choice of the best logistic model by a composite score.

<table>
<thead>
<tr>
<th>Fatty acids</th>
<th>Odds ratios</th>
<th>95% confidence interval</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linoleic</td>
<td>0.81</td>
<td>0.67–0.98</td>
<td>0.03</td>
</tr>
<tr>
<td>Arachidonic</td>
<td>3.19</td>
<td>1.48–6.87</td>
<td>0.003</td>
</tr>
<tr>
<td>DHA</td>
<td>0.15</td>
<td>0.04–0.56</td>
<td>0.004</td>
</tr>
</tbody>
</table>

3. Results

After having checked with the Shapiro-Wilk test, all of the 12 FA followed normal distribution: myristic: P = 0.07, palmitic: P = 0.22, palmitoleic: P = 0.10, stearic: P = 0.09, oleic: P = 0.79, linoleic: P = 0.85, gammalinolenic: P = 0.43, DHGL: P = 0.07, arachidonic (AA): P = 0.61, alphalinolenic: P = 0.08, eicosapentaenoic acid (EPA): P = 0.18, and docosahexaenoic acid (DHA): P = 0.64. Simple statistics in univariate analysis are shown in Table 1. Four FA had significantly different levels between the two groups: stearic acid, AA, EPA, and DHA.

Relative differences between group 1 and group 2 are displayed in Figure 1. Positive differences were associated with higher values in group 2, while negative differences were associated with higher values in group 1. The highest positive relative differences were for stearic (11.6%), gammalinolenic (11.4%), and arachidonic (11.2%) acids. The lowest values for negative relative differences were represented by EPA (−33.9%), DHGL (−25.0%), and linoleic acid (−3.8%). Eight FA out of 12 were higher in patients with periodontitis, while 4 FA were higher in the control group.

The Hosmer-Lemeshow test showed that the relationship between the FA levels and the response variable was well fitted by the logistic regression model: chisquare = 10.43, with a P value of 0.24. The mean VIF between predictors was equal to 2.19, with a maximum value of 3.12, which means that the collinearity between the FA is low. Then, ROC curves were constructed for each FA. They are displayed in Figure 2. The AUC calculation showed very low prediction performance values for each FA separately. The construction of a composite score allowed choosing the best logistic model (Table 2). The latter was obtained with these 3 FA: arachidonic, linoleic, and DHA. These FA, which are finally retained by the model, are highlighted in Figure 1 (hatched bars). Another ROC curve was yielded by means of this composite score (Figure 3). AUC was found to be equal to 0.821. Depending on the threshold value, the best results, after cross-validation, were 26 true positives, 18 true negatives, 8 false positives, and 4 false negatives. So 44 subjects were correctly classified and 12 were misclassified. Sensitivity, specificity, positive predictive value and negative predictive value, were 86.7%, 69.2%, 76.5%, and 81.8%, respectively (Table 3). The quality of prediction was compared in Table 3. Sensitivity and specificity did not significantly differ between the model yielded by means of the composite score versus the full model including 12 covariates: 86.7% versus 80.0% (P = 0.46) and 69.2% versus 73.1% (P = 0.73), respectively.

4. Discussion

4.1. Fatty Acids Levels and Periodontal Disease. The FA levels in serum differed significantly depending on the presence of periodontitis (moderate or severe according to Page and Eke [10]). The present study was aimed to check if inflammation that was clinically detected in periodontal tissues could be...
linked with biological measurements, such as the FA levels, which may include different forms of periodontitis. Discrimination between those forms was not investigated. The three FA selected by the model are playing a significant role in bone metabolism [12–14]. Firstly, AA was found to be more abundant among diseased patients (Table 1 and Figure 1), which confirms the results from other studies [8, 15] and its implication in periodontal inflammation. It was also well documented by Eberhard et al. [16] in experimental gingivitis and by Figueredo et al. [17] who found a significant relation between the serum level of FA and the severity of periodontitis. But the role of linoleic acid (18:2n-6), an essential polyunsaturated FA, which appeared to be one of the most discriminant variables, needs to be more investigated. Unlike AA, it was found to be less abundant among diseased patients. Johnson and Fritsche [18] found that linoleic acid reduced the risk of some diseases, but at a higher level it might contribute to excess chronic inflammation. The roles of EPA and DHA, which are ω-3 fatty acids, were also found to be prominent. Their role is also described in the literature: patients with a low DHA or EPA intake are more likely to have periodontal disease [19, 20]. They also regulate hepatic lipid and glucose metabolism [21].

Since this study follows a cross-sectional design, it is difficult to interpret the variations of fatty-acids in blood, which are known to change over time. So one can imagine little variations of concentration and inconstant results about the significant effect of some FA. However, the results of the present study are in agreement with longitudinal studies [22]. It is now well known that a competition between n-6 and n-3 unsaturated FA occurs in prostaglandin formation [19], which mainly implies AA, EPA, and DHA. Another assumption could be that the variation of the FA concentration in blood may precede the clinical evidence and the oral symptoms.

### 4.2. Variable Selection Method.
Usually, reducing the number of predictors leads to a reduction of the discrimination power [23, 24]. However, the present method allowed reducing the number of variables from 12 to 3, with no reduction of power of discrimination between the two groups or patients and with no reduction of the number of patients correctly classified, when cross-validation procedure was applied. One main finding is that the ordinary logistic model, including the variables which were significant in univariate analysis, yielded lower values of sensitivity and specificity than the logistic model with composite score (Table 3).

It is worthwhile to underline that the FA retained by this selection method were not systematically the most prominent in univariate analyses when discriminating the 2 groups of patients (Table 1). In univariate analyses, stearic acid,
AA, EPA, and DHA showed significantly different levels between cases and controls. But the variable selection method proposed in this paper showed that the subset of variables the most relevant was constituted by AA, linoleic acid, and DHA (Figure 1). The outcomes of the multivariate model do not systematically match with those of the univariate models, because the multivariate approach takes into account the best combination between the 12 variables to explain the "group" difference. The cross-validation procedure may partly explain those differences, because the performance of our model was less penalized by it. However, cross-validation is a more objective approach, by testing the model on a different subset of subjects: its goal is to gauge the generalizability of the prediction model. Another problem may rise from collinearity or correlation between explanatory variables in multiple regression, which can lead to unexpected results [24]. With the model proposed in the present study, the mean VIF between predictors was equal to 2.19, with a maximum value of 3.12, which is much lower than 10 [25]. So, absence of multicollinearity could be assumed.

FA levels in serum of patients were significantly different according to the presence of periodontitis. Of course, FA are not the only components in these complex biological processes, but their importance has been demonstrated. In such a multifactorial disease, it is uncertain to assign a patient to one of the two groups simply by relying on the presence of a single risk factor [26]. Nevertheless, diagnosis by FA measurements is expected to be made earlier: blood measurement data, which result from biochemical reactions during complex pathological processes, are indeed expected to precede the clinical symptoms. In a way, this could help to anticipate the clinical symptoms. This study aimed to help the practitioner in making clinical decisions, such as diagnosis, treatment plan, and prognosis, but it does not have to

**Figure 2:** ROC curves for the 12 fatty acids.
be considered as a perfect rule which should replace the clinician’s experience.

More generally, by taking into account the comparison of ROC curves, our approach could optimize the choice of variables in multivariate analyses and could better fit in with prognosis of oral diseases in medical research. Very few tools are available in epidemiological research to best choose a multivariate model when many explanatory variables have been measured and are potentially relevant. So, in the future, it will be interesting to conduct a follow-up study in order to understand whether the biochemical transformations, followed by the FA measurements, could precede the clinical and pathological manifestations. This could lead to an earlier diagnosis and a more efficient prophylactic intervention, because taking into account the biochemical measurements on an asymptomatic subject is an act of prevention [27].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[14] L.-S. Kremmyda, E. Tvrzicka, B. Stankova, and A. Zak, "Fatty acids as biocompounds: their role in human metabolism, health and disease—a review—part 2: fatty acid physiological roles and


