STATISTICAL ANALYSIS IN ORAL LICHEN PLANUS

N. TEODORESCU[1]
A. PETRESCU[2]
M. COSTACHE[3]
Ş. ŢOVARU[4]
[1] Technical University of Civil Engineering, Bucharest, Romania
[2] INCD "Victor Babeş", Bucharest, Romania
[3] "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania
[4] "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania

***Abstract**: The aim of this article is to present a statistical analysis of clinical parameters of patients with oral lichen planus (OLP). The study was performed on real data which included 92 patients. We intend to present the correlation between various clinical parameters in order to establish the most important factor which can influence the medical decision. The Model Selection Loglinear Analysis procedure is used to identify models for describing the relationship between variables. This is accomplished through analysis of the cell counts of the crosstabulation table formed by the cross-classification of the interested variables.*

***Keywords:** correlation, loglinear analysis, oral lichen planus.*

INTRODUCTION

Oral lichen planus (OLP) is a subacute or chronic disease, a common immunologically mediated condition of the mucosa, appearing mostly idiopathic. The lesions result from a complex interaction between cellular and molecular signals that begin with alteration of basal keratinocyte antigens by an endogenous or exogenous agent. Histopathologically, OLP demonstrates parakeratosis or even hyperorthokeratosis; the epithelium is attenuated or eroded in the atrophic type and hyperplastic in the reticular type. There are variable numbers of colloid (Civatte) bodies. Saw-tooth rete ridges can be seen and the basal cells are degenerated. A band-like lympho-histiocytic infiltrate hugs the connective tissue - epithelium interface. Also, melanophages can be present in the lamina propria.

In the present study, the authors analyzed 92 patients with various oral lesions.

MATHEMATICAL MODEL

Descriptive statistics for the variables analyzed were calculated. In the statistical analysis, Spearman correlation coefficient and the model selection loglinear analysis procedure were used. In all cases, the following data was recorded: age, sex, oral location, clinical appearance, and histopathological parameters.

Until the late 1960's, contingency tables - two-way tables formed by cross classifying categorical variables - were typically analyzed by calculating chi-square values and testing the hypothesis of independence. When tables consist of more than two variables, researchers compute the chi-squares for two-way tables and then again for multiple sub-tables formed from them in order to determine if associations and/or interactions had taken place among the variables. In the 1970's the analysis of crossclassified data changed quite dramatically with the publication of a series of papers on loglinear models by L.A. Goodman. Many other books appeared around that time builded on Goodman's work (Bishop, Fienberg & Holland 1975 [1]). Now researchers were introduced to a wide variety of models that could be fitted to crossclassified data [2], [3], [5]. Thus, the introduction of the loglinear model provided them with a formal and rigorous method for selecting a model or models for describing associations between variables. The basic strategy in loglinear modeling involves fitting models to the observed frequencies in the cross-tabulation of categorical variables. The models can then be represented by a set of expected frequencies that may or may not resemble the observed frequencies. Models will vary in terms of the marginals they fit, and can be described in terms of the constraints they place on the associations or interactions that are present in the data. The pattern of association among variables can be described by a set of odds and by one or more odds ratios derived from them. Once expected frequencies are obtained, we then compare models that are hierarchical to one another and choose a preferred model, which is the most parsimonious model that fits the data. It's important to note that a model is not chosen if it bears no resemblance to the observed data. The choice of a preferred model is typically based on a formal comparison of goodness-of-fit statistics associated with models that are related hierarchically (models containing higher order terms also implicitly include all lower order terms). Ultimately, the preferred model should distinguish between the pattern of the variables in the data and sampling variability, thus providing a defensible interpretation.

The following model refers to the traditional chi-square test where two variables, each with two levels (2 x 2 table), are evaluated to see if an association exists between the variables.

$$Ln(F_{ij}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

$Ln(F_{ij}) =$ the log of the expected cell frequency of the cases for cell noted ij in the contingency table.

$\mu =$ the overall mean of the natural log of the expected frequencies.

$\lambda =$ term in which each represents "effects" of the variables on the cell frequencies.

A and *B* = the variables

i and *j* = refer to the categories within the variables.

Therefore

$\lambda_i^A =$ the main effect for variable *A*.

$\lambda_j^B =$ the main effect for variable *B*.

$\lambda_{ij}^{AB} =$ the interaction effect for variables *A* and *B*.

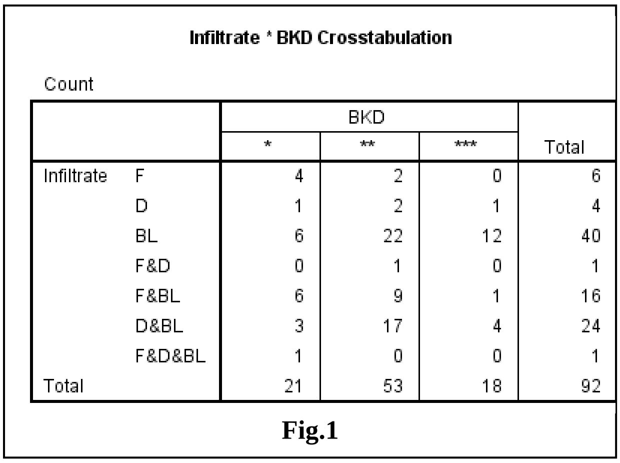
The above model is considered a Saturated Model because it includes all possible one-way and two-way effects. Given that the Saturated Model has the same amount of cells in the contingency table as it has effects, the expected cell frequencies will always match exactly the observed frequencies, with no degrees of freedom remaining [4]. For example, in a 2 x 2 table there are four cells and in a saturated model involving two variables there are four effects, μ, λ_i^A, λ_j^B, λ_{ij}^{AB}, therefore the expected cell frequencies will match exactly the observed frequencies. Thus, in order to find a more parsimonious model that will isolate the effects which best demonstrates the data patterns, a non-saturated model must be sought. This can be achieved by setting some of the effect parameters to zero. For instance, if we set the effects parameter lij AB to zero (i.e. we assume that variable A has no effect on variable B, or vice versa) we are left with the unsaturated model.

The aim is to construct a model in such way that the cell frequencies in a contingency table are accounted using the minimum number of terms. This is done by a process of backward elimination. What this means is that one begins with the maximum number of terms, and then drops a term in each round. Statisticians refer to it as the backward hierarchical method. In practice, one commences the analysis by including all the variables. This is referred to as the saturated model. It can usually be expected to predict the cell frequencies perfectly. Then the highest order interaction is removed, and its effect on how closely the model can now predict the cell frequencies is noted. This process of progressive elimination is continued. Each time a variable is removed a statistical test is performed to determine whether the accuracy of prediction falls to such an extent that the component most recently eliminated could be one of the components of the final model. At each stage the assessment of goodness-of-fit is made by means of a statistical test known as the likelihood ratio. The final model includes only the associations necessary to reproduce the observed frequencies. A comparison of the observed and expected frequencies for each cell using the likelihood ratio makes the evaluation of the final model. In the same way as in the case of χ^2 test, small expected frequencies can lead to loss of power. It is recommended that all expected frequencies should be greater than 1, and no more than 20% should be less than 5.

The basic strategy in loglinear modeling involves fitting models to the observed frequencies in the cross-tabulation of categorical variables. The models can then be represented by a set of expected frequencies that may or may not resemble the observed frequencies. Models will vary in terms of the marginals they fit, and can be described in terms of the constraints they place on the associations or interactions that are present in the data. The pattern of association among variables can be described by a set of odds and by one or more odds ratios derived from them. Once the expected frequencies are obtained, we then compare the models that are hierarchical to one another and choose a preferred model, which will represent the most parsimonious model that fits the data. It's important to note that a model will not be chosen if it bears no resemblance to the observed data. The choice of a preferred model is typically based on a formal comparison of goodness-of-fit statistics associated with other models that are related hierarchically (models containing higher order terms also include all lower order terms by default). Ultimately, the preferred model should distinguish between the pattern of the variables in the data and he sampling variability, thus providing a defensible interpretation

RESULTS AND DISCUSSIONS

For our sample, we observed that atrophy of the epithelium was absent in 56.5% (52cases) and present in 43.5% (40cases), acanthosis was absent in 39.1% and was present in 60.9%. Most patients (77.2%) were presented with parakeratosis. In only 6.5% of the cases ortokeratosis could be observed and 16.3% of the cases presented the keratosis form. Regarding the lymphocytic infiltrate, we noticed that both lymphocytes and macrophages are present in 38 patients. Only 20 cases presented lymphocytes, macrophages and plasma cells and 34 cases present only lymphocytes. Fig. 1 shows a crosstabulation for vacuolar change or destruction of basal keratinocytes (BKD) and inflammatory infiltrate (classified as cellular type). Most of the patients (22 cases) had lymphocytic infiltrate arranged in a band-like pattern and medium- grade BKD (**).



Infiltrate * BKD Crosstabulation

Count

		BKD			Total
		*	**	***	
Infiltrate	F	4	2	0	6
	D	1	2	1	4
	BL	6	22	12	40
	F&D	0	1	0	1
	F&BL	6	9	1	16
	D&BL	3	17	4	24
	F&D&BL	1	0	0	1
Total		21	53	18	92

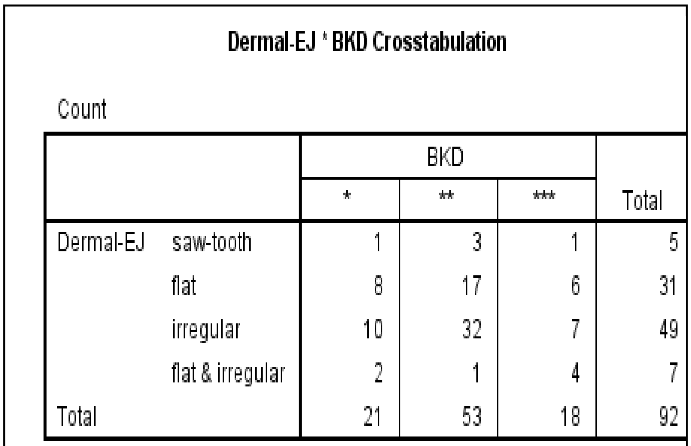
Fig.1

Spearman correlation coefficient was used. A high inverse correlation between the atrophic form and acanthosis (rho=-0.914, p<0.01) was observed. Moreover, Dermal-EJ aspect was statistically correlated with the atrophic form (rho=-0.503, p<0.01) and the presence of acanthosis (rho=0.553, p<0.01).

In clinical investigations we often have responses and explanatory variables that are both categorical. A very good mathematical tool that fits our data is loglinear analysis.

We used atrophy, acanthosis and basal keratinocytes destruction in order to find an appropriate model. These variables fulfill all the conditions for applying loglinear analysis.

Medium-grade BKD (**) with irregular dermal-epithelial junction (Dermal-EJ) was the most frequent found characteristic (32 cases), whereas saw-tooth epithelial junction aspect was observed in only 5 cases (Fig.2).

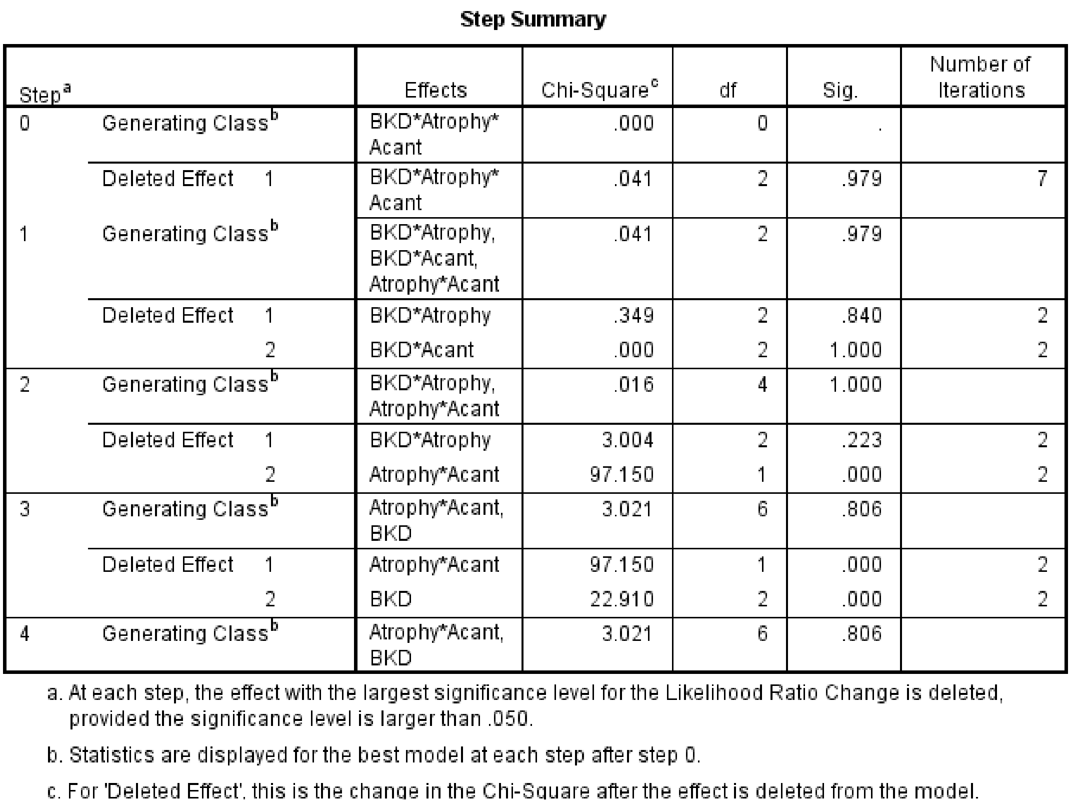


Dermal-EJ * BKD Crosstabulation

Count

		BKD			Total
		*	**	***	
Dermal-EJ	saw-tooth	1	3	1	5
	flat	8	17	6	31
	irregular	10	32	7	49
	flat & irregular	2	1	4	7
Total		21	53	18	92

Fig.2



Step Summary

Step[a]			Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]		BKD*Atrophy*Acant	.000	0	.	
	Deleted Effect	1	BKD*Atrophy*Acant	.041	2	.979	7
1	Generating Class[b]		BKD*Atrophy, BKD*Acant, Atrophy*Acant	.041	2	.979	
	Deleted Effect	1	BKD*Atrophy	.349	2	.840	2
		2	BKD*Acant	.000	2	1.000	2
2	Generating Class[b]		BKD*Atrophy, Atrophy*Acant	.016	4	1.000	
	Deleted Effect	1	BKD*Atrophy	3.004	2	.223	2
		2	Atrophy*Acant	97.150	1	.000	2
3	Generating Class[b]		Atrophy*Acant, BKD	3.021	6	.806	
	Deleted Effect	1	Atrophy*Acant	97.150	1	.000	2
		2	BKD	22.910	2	.000	2
4	Generating Class[b]		Atrophy*Acant, BKD	3.021	6	.806	

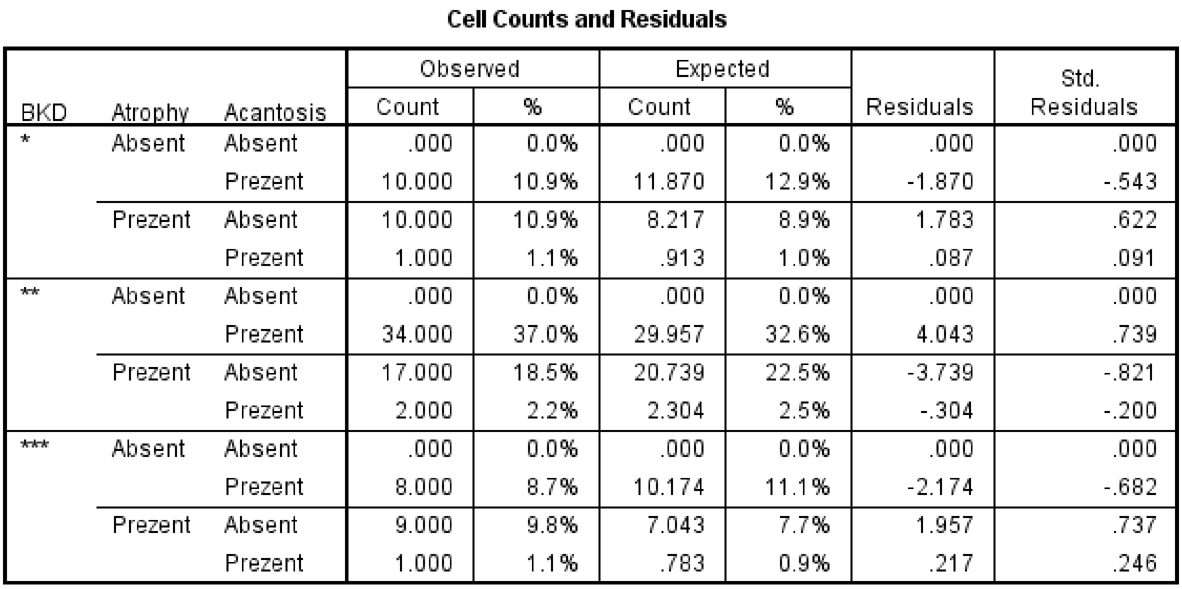
a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Fig. 3

A model in which basal keratinocyte destruction was considered the main effect and atrophy/acanthosis (*) the interaction effect was obtain using backward elimination (Fig. 3, Fig 4).

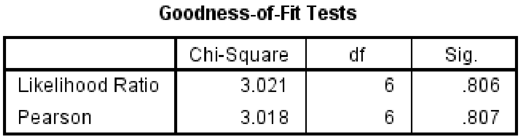


Cell Counts and Residuals

			Observed		Expected			
BKD	Atrophy	Acantosis	Count	%	Count	%	Residuals	Std. Residuals
*	Absent	Absent	.000	0.0%	.000	0.0%	.000	.000
		Prezent	10.000	10.9%	11.870	12.9%	-1.870	-.543
	Prezent	Absent	10.000	10.9%	8.217	8.9%	1.783	.622
		Prezent	1.000	1.1%	.913	1.0%	.087	.091
**	Absent	Absent	.000	0.0%	.000	0.0%	.000	.000
		Prezent	34.000	37.0%	29.957	32.6%	4.043	.739
	Prezent	Absent	17.000	18.5%	20.739	22.5%	-3.739	-.821
		Prezent	2.000	2.2%	2.304	2.5%	-.304	-.200
***	Absent	Absent	.000	0.0%	.000	0.0%	.000	.000
		Prezent	8.000	8.7%	10.174	11.1%	-2.174	-.682
	Prezent	Absent	9.000	9.8%	7.043	7.7%	1.957	.737
		Prezent	1.000	1.1%	.783	0.9%	.217	.246

Fig. 4

The result was $\chi^2(6) = 3.021$ with a p=0.806. These values show a high degree of fit between the observed frequencies and expected frequencies generated by the model (Fig. 5).



Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	3.021	6	.806
Pearson	3.018	6	.807

FIG. 5

CONCLUSION

We found in our study a good model for describing OLP lesions. Using backward elimination, we demonstrated that basal keratinocytes destruction is the main effect and atrophy/acantosis represents the interaction effect. Our values, $\chi^2(6) = 3.021$ and p=0.806, indicated that we have a high degree of fit between the observed frequencies and expected frequencies generated by this model. In perspective, we hope to discover other models, using more variables. Other important parameters which we intend to study with linear analysis are the infiltrate (cellular) type and keratotic form (parakeratosis, ortokeratosis).

BIBLIOGRAPHY:

[1] Bishop, M M, Fienberg S E and Holland P W, *Discrete multivariate analysis: theory and practice*. Cambridge. MA: MIT Press, 1975.

[2] Garson, G. D.,:*Log-Linear Analysis*. Asheboro, NC: Statistical Associates Publishers, 2012.

[3] Howell, D., *Statistical Methods for Psychology*, Cengage Learning, 2013

[4] Knoke, D. and P.J. Burke, *Log-Linear Models*. Sage Publications, Inc. Newberry Park, California, USA, 1980.

[5] Tabachnick, B., Fidell, L.:*Using Multivariate Statistics (6th Edition)*, Pearson Education Publisher, 2012.

.