SÃO PAULO MEDICAL JOURNAL
LETTER

Methodological aspects of prognostic factor studies: some caveats

To the Editor: In this comment some methodological aspects of prognostic factor studies are discussed. Categorization of continuous variables should be avoided and when necessary must be done before applying a test. Confidence intervals give more precise and clinically relevant information than p-values. Statistical tests providing a maximum of test power should be chosen. Repeated applications of the Cox model after bootstrap resampling permit an estimation of the stability of the model.

Recently Souto and co-workers[1] created a new scoring system for patients with myelodysplastic syndromes.We would like to discuss some methodological items of this study.

Continuous variables were categorized by trial and error until p-values were found close to 5%. It may be useful to divide a continuous variable into categories, but this is dangerous because categorizing means to a certain degree "throwing away information", or, if done by a data-dependent approach, may increase substantially the type I error rate. A better approach would be the use of univariate Cox models for the continuous variables.[2]

The authors stress the importance of the p-value, e.g. using the phrase " the most significant prognostic factor" P-values are only used to test whether the null hypothesis should be rejected, thus indicating whether there is any difference at all, but are never an estimation of the amount of the difference. A minimal difference may be statistically significant but clinically without any importance. Therefore confidence intervals of beta values of the prognostic factors would have provided more precise information than p-values.

According to Sanz's work,[3] the survival curves of the variable hemoglobin were divided into three groups, but no significant difference could be found, whereas this was shown after dichotomization. We should not forget that Sanz et al.[3] analyzed a total of 370 patients, which gives much more statistical power for detecting real differences between the groups. Furthermore, when comparing the survival curves of ordered groups (here: hemoglobin), trend tests must be applied. Otherwise the statistical power would be reduced, such that real associations could be missed.

The final Cox model depends on the variables initially offered, the rules determining the inclusion and removal of variables, and the population selected. Since in most statistical programs the user can choose between several possibilities (e.g. between conditional parameter estimates, maximal partial likelihood estimates, Wald statistics, or between foreward or backward selection etc.) the calculation may result in different models. In stepwise Cox regression the variables selected may be highly unstable and other samples from the same population might result in different models. A model formulated and fitted to the same data, as in this case, is usually "overfitted". Under ideal conditions the model should be created by a "learning set" of data and than tested with the data of other patients ("test set"), but this requires a large number of patients. Alternatively, repeated calculations of the Cox model based on a large number of data sets obtained by bootstrap resampling of the original 59 patients would show the stability and predictive ability of the model.[4] A study based on 59 patients has a low test power and its results should at best be regarded as hypothesis forming.

We do not understand why the variables E/M ratio and WBC were included in the scoring system, since neither of them had entered the multivariate Cox model and therefore must be considered to be of no detectable prognostic relevance. Moreover, we do not understand why the authors divided the scoring system into exactly three risk groups. Was this number suggested, for instance, by a histogram or by a cluster analysis?

The above mentioned items should be clarified before we can accept the new prognostic scoring system.

Konradin Metze, PhD
Universidade Estadual de Campinas, São Paulo, Brazil.

References

1. Souto EX, Chauffaile MLLF, Moncau JEC et al. Myelodysplastic syndromes (MDS): prognostic factors and scoring system. Rev Paul Med 1997;115(5):1537-41
2. Metze K. Methodological problems of grading tumour regression: responders compared to non-responders. J Cancer Res Clin Oncol 1998;124:281-2.
3. Sanz GF, Sanz MA, Vallespi T, et al. Two regression models and a scoring system for predicting survival and planning treatment in myelodysplastic syndromes: a multivariate analysis of prognostic factors in 370 patients. Blood 1989;74(suppl 1):395-408.
4. Altman DG, Andersen PK. Bootstrap investigation of the stability of a cox-regression model. Stat Med 1989;8:771-83.

In response: We would like to express our thanks for the critical reading and valuable comments.

We do not agree with the point of view of the author of these criticisms against the categorization of continuous variables. Both the analysis using categorized variables and the analysis using continuous variables encompass the use of modeling for data representation. The model will be good or not depending on how good it represents the relationship under study. To categorize a continuous variable does not imply necessarily losing important information, and the use of more complex models does not guarantee automatically a better representation of data. The analysis of tabular data is an indispensable preliminary to modeling and can also serve as a cross-check on modeling results.[1] The procedure we used can cause a significant increase in probability of type 1 error, and we agree that it has to be avoided, but not for this reason. This procedure should not be used because it can introduce bias in the estimates, as it is done using this effect as the criterion for choosing the category boundaries. There are other ways to choose the cutpoints without introducing bias. Regarding the emphasis on the use of the p-value we believe the author of the criticisms has mixed up statements we made in the paper with statistical concepts that have no relation to our work. We agree that the use of confidence intervals to present results is more informative than the use of p-values and should be used often, following the trend of abandoning the dichotomization imposed by the use of hypothesis testing, in scientific inference. However his suggestion that we used p-values ad should be used often, following a trend to abandon the dichotomization imposed by the use of hypothesis testing in scientific inference. However his suggestion that we used p-values as a measure of effect size is not true. Just after the statement he cites, "most significant prognostic factor", we presented the data with the description of the observed differences in survival.

We also agree that a trend test should be used every time we analyze ordered categorized variables, if the test is available. And that was not the case when we did the analysis. After the publication of the paper, the test was done with no significant statistical differences.

Two score systems were developed, one based on univariate analysis and the other based on multivariate analysis, with agreement in the results. Because of space limitation we chose to present only the table with results based on the univariate analysis. The subjects were split into three groups according to a histogram of the risk.

Any data set can generate a great number of different analyses, depending on the degree of utilization of the information in the sample, on the procedures the statistician decides to use, and on other subjective decisions like the degree of complexity of the analysis. In this case we decide not to use couples procedures as the bootstrap.

We understand that the validity of our work is not affected by the criticisms. Any scientific work, even when using complex models in the analysis or a large sample size, cannot have its results evaluated in an "accept or reject" way. Any results can be potentially explained by sample variability, model misspecification, misclassification, bias, or mistakability, and they can also correspond to a real association. Within the conditions presented in the article and in these explanations, the results should be weighed up and used as one more contribution to the discussion leading to criteria for the prognostic evaluation of MDS patients. We believe that additional work in this field is neccessary in order to achieve reliable score systems.

Elizabeth Xisto Souto
Faculdade de Medicina de Botucatu, Escola Paulista de Medicina/UNIFESP, São Paulo, Brazil.

Reference

1. Greenland S. Problems in the average-risk interpretation of categorical dose-response analyses. Epidemiology 1995;6: 563-565.