Reviewer's report

Title: Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung

Version: 3
Date: 24 February 2014
Reviewer: José Caldas

Reviewer's report:

The authors propose a statistical framework to discriminate between high and low-survival groups among patients diagnosed with stage I or II squamous cell carcinoma of the lung. My main comments concern the broad applicability of the proposed framework and the intrinsic merits of performing binary classification vs. survival analysis. Specific comments to the authors follow below.

Major Compulsory Revisions

- I believe one of the main points in the paper is that, in some data sets, the large number of right-censored patients and the specificities of the pathology (e.g. co-morbidity) make it less useful to estimate covariates in a survival model (e.g. the standard Cox proportional hazards model) than to simply partition patients into low and high survival groups and perform binary classification. However, in section "Logistic regression versus Kaplan Meier analysis", the authors compare the methods on the basis of which genes were found by each method (e.g., "Although our KM analysis did identify some immune-related genes as prognostic, the logistic regression approach proved superior in that it identified a much larger number of highly correlated B cell genes in the stage I and II cases of the GDS2373 data set.")., which does not seem entirely convincing. Would it be possible to perform a performance-based comparison, for instance based on the cross-validation procedure described in the paper?

- Still related to the first comment, if we performed survival analysis on the same patient subset as the one used for performing logistic regression, then wouldn't the right-censoring problem be avoided and wouldn't covariates obtained via the Cox proportional hazards model be more informative than the covariates obtained via logistic regression, since the former model takes into account survival times?

- "The logistic regression approach is less affected by incomplete or heavily right-censored survival data than KM analysis". If we were to add more patients to the low and high survival groups, wouldn't we also end up having many right-censored cases in the logistic regression case? In general, isn't this binary classification setting doomed to include only a small proportion of the patient
sample, in order to avoid including multiple right-censored cases? Although the authors state "Moreover, the intermediate survival cases, which are heavily right censored and may thus degrade the analysis, are of lesser significance for predicting survival class and need not be used", would this hold in general?

- the method includes a number of steps which are fine-tuned to the analyzed data set (e.g. gene filtering described at the by interquartile criterion on page 6; exclusion of four patients ). Additionally, only a single data set was used to test the proposed method and no comparisons to other binary classification methods were made. How hard would it be to make the proposed approach broadly applicable to survival microarray data sets?

Minor Essential Revisions

- The paper would benefit greatly from a figure describing the workflow, as it would be far easier to follow the many steps involved in the analysis.

Discretionary Revisions

- "one-left-out" ---- the standard name for this validation approach is "leave-one-out"
- pages 6/7, "selectivity" ---- the standard name for the formula TN/(TN+FP) is "specificity".

Level of interest: An article of limited interest

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests