Author's response to reviews

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 and triple negative breast carcinoma

Authors:

David L Mount (mount@email.arizona.edu)
Charles W Putnam (cwp@u.arizona.edu)
Sara M Centouri (smb4@email.arizona.edu)
Ann M Manziello (annm@email.arizona.edu)
Ritu Pandey (ritu@email.arizona.edu)
Linda L Garland (lgarland@email.arizona.edu)
Jesse D Martinez (jmartinez@azcc.arizona.edu)

Version: 7 Date: 2 May 2014

Author's response to reviews: see over
April 28, 2014

BMC, Medical Genomics

Dear Editor,

Please find uploaded onto your web site the revised version of our manuscript titled “Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma”. In addition, we have also uploaded a separate document that contains our responses to the reviewer’s comments. As you will see from our revised manuscript and from our responses to the reviewers we have addressed all of the concerns that were raised during the previous review. Importantly, we were able to validate our approach and the gene signature we identified by using a data set of triple negative breast cancers. Because this added analysis of breast cancers was consistent with our lung cancer analysis we elected to draw attention to this by including this in the title of the manuscript.

Please do not hesitate to contact me should you have questions regarding our manuscript.

Sincerely,

Jesse Martinez, Ph.D.
Professor, Cell & Molecular Medicine
University of Arizona Cancer Center
RESPONSE TO REVIEWER COMMENTS AND QUERIES

We appreciate the thoughtful, insightful comments and queries from both reviewers and agree with many of them. We address their concerns here and have made corresponding revisions and additions to the manuscript. In particular, we have analyzed a triple negative breast cancer data set using the same approaches and have demonstrated the role of immune system-related genes in predicting tumor recurrence in that malignancy as well.

We have put a great deal of effort into finding and analyzing a suitable second data set and believe that the revised manuscript is greatly improved thereby and hope that it is now acceptable for publication.

1st Reviewer (Caldas)

As noted by the reviewer, in our manuscript we make the argument that the observation that most of the genes predictive of long survival (versus early death) are correlated with each other provides evidence that our methodology did indeed identify predictive genes representing a common process (B-cell infiltration). By way of contrast, our Kaplan Meier analysis neither ranked the B-cell genes so highly nor included so many amongst the “best predictive genes”. For a more “performance-based comparison,” as suggested by the reviewer, we have analyzed the reduced set of 40 cases and found that a subset of the same immune system genes remain predictive for survival despite even though the data set had been reduced in size.

It was not possible however to do a comparable validation analysis, as suggested by the reviewer, as the KM method predicts survivals based on expression quantiles. We argue that finding immune-system genes using both methods serves as a further validation check. We are, presently investigating these results in further laboratory and clinical studies as we believe that they have highly significant implications for the treatment of lung and breast cancer.

The reviewer is correct in stating that survival analysis based on the same patient subsets as we used for logistic regression analysis would similarly avoid the right-censoring problem but that argument does not address our concern that precise duration of survival in the early-death group is not directly representative of disease progression. As we state on page 9, “...in the elderly population typical of SQCC, patient deaths not infrequently result from co-morbidities [43], such as infection, heart disease, stroke, emphysema and diabetes, rather than cancer. Duration of survival, as in the KM method, is therefore an inadequate proxy for disease progression.” We suggest, based upon our analysis, that treating the earliest deaths as a unitary group – that is, ignoring precisely when they died – avoids making the perhaps untenable assumption, namely: that within that group of patients, the time of (early) death is directly correlated with disease progression.

Some of the patients who were members of the long-survivor group had right-censored survival data but obviously that is of little concern – they could only have lived longer than estimated! In a data set such as this one, which has good follow-up data, early death cases should not be (and were not) right censored. Of the cases falling in between (65 patients), about 48 were right censored. Even so, using our two groups of 40 patients, (80 of 107 cases), only one early death case and only 5 in the long survival group were in fact right-censored. In the end, the quality of the analysis depends heavily upon the quality of the clinical data, especially the proportion of cases with long follow-ups and accurate survival data. With this particular data set, we were able to include 80 (75%) of 107 cases, in no small part because of the quality of the clinical data.
Gene filtering by the interquartile criterion was chosen to limit the analysis to genes that are varying the most in the clinical samples, and to avoid instances of small, less significant variations that are in greater doubt.

The four patients excluded from the leave-one-out analysis all died within two months of operation. In our clinical experience (authors CWP, a surgeon, and LLG, a medical oncologist), early stage SQCC patients who succumb early after operation are far more likely to do so from postoperative complications, such as infection, technical misadventures, pulmonary insufficiency, etc. than from recurrence of malignancy. Including those cases would only weaken the analysis.

A figure showing the work flow has been added and the terms “leave-one-out” and “specificity” have been inserted where appropriate.

2nd Reviewer (Sun)

We have added to the manuscript a statement comparing our approach to the Cox proportional hazard model (p. 19).

We have also included in the revised manuscript an analysis of a second data set using our approach (see p. 19-210.

We whole-heartedly concur with Dr. Sun’s comment that “lung cancer patient survival is rather complex and contributed (to) by multiple factors. Tumor molecular signature is just part of (the) story.” In our analysis, we have attempted to deal with that complexity in several ways, as summarized here.

First, consonant with the objective of the study, we limited case selection to stages I and II.

Second, we chose two starkly contrasting patient groups: earliest deaths and long-term survivors.

Third, in our comparisons we used groups of equal size, thereby avoiding model bias. Furthermore, a group size of 20 is suitable for analysis of a single variable (Hosmer) but is not satisfactory for analyzing multiple variables.

Fourth, we limited the use of right-censored data among the “earliest deaths” group (in the longest survivors group, right-censoring has minimal impact).

Finally, we excluded four deaths occurring very soon after operation, mortality almost certainly ascribable to factors (infection, respiratory insufficiency, operative complications) other than recurrence of SQCC.

Within those constraints, we nonetheless were able to identify 24 gene expressions (most of which were highly correlated B-cell related genes) highly predictive of early death versus long-term survival. We suggest that these observations support the validity of our methodology, as well as adding SQCC of the lung to the growing list (perhaps also including triple-negative breast cancer) of solid tumors in which higher levels of B-cell gene expressions indicate an increased likelihood of prolonged survival.

Editor’s Summary

Emphasizing criticisms common to both reviewer’s, the editor requested:
1. A more thorough comparison with conventional risk modeling methods.
2. A more complete validation study.

As described here, we have attempted to conscientiously and thoroughly address both issues. We add here that, as suggested in 2, we have performed an additional validation analysis using a multiple leave out analysis as recommended. The results are given in the validation section and indicate a high level of accuracy for the two most predictive genes for long survival in the lung study.