Reviewer's report

**Title:** Clinical and multiple gene expression variables in survival analysis of breast cancer: Analysis with the hypertabastic survival model

**Version:** 1  **Date:** 22 March 2012

**Reviewer:** Khalil Helou

**Reviewer's report:**

With the overwhelming preponderance of the Cox–regression in the medical literature it is nice to be reminded once in a while of the alternative parametric models. This manuscript represents a rather interesting attempt by the authors. However, I can’t help wishing for a shorter, better structured and leaner manuscript. The manuscript is very hard to read in its present form. Even after re-reading the paper several times I struggled to fully comprehend it.

As I understand the purpose of the paper is two-fold. First, promoting the hypertabastic survival model; second, by a multivariate modeling approach do give insights into time dependent progression of the disease.

As a general impression the case for the hypertabastic distribution is fair, however it’s not convincing enough. The previous article by the authors on the subject, is much more convincing (Theoretical Biology and Medical Modelling, 2007). Moreover, key elements are left out which makes its utility somewhat limited. On the other hand, the multivariate modeling seems a bit redundant. The data set the authors use it’s a ‘classical’ one, it was used as example in other methodical studies as well. The fact that multivariate modeling is useful does not need yet another proof.

If the main aim of the manuscript is to introduce the hypertabastic survival regression then I would suggest the authors to leave the extensive modeling work out, and focus on the method itself. If they wish to keep the modeling part, then it was to be shortened down and to be more on focus.

**Major Compulsory Revisions**

1) I think the manuscript could benefit form a better structuring.

2) Both in the abstract and conclusion the authors claim that through the use of the hypertabastic model they attained an accuracy superior to the Cox-regression. However, it’s not clear how they arrived to this conclusion. As I see accuracy denotes the closeness of the estimates to the exact or true values. As no simulation study was conducted this cannot be assessed. I guess due to the rather complicated pdf of the hypertabastic distribution there are no closed solutions for simulating survival times.

3) Introduction: The introduction needs to be scaled down and to be more on focus. The main questions and aims need a more clear formulation.

4) Methods section: given the readership of the journal, the authors have to make
a better point describing the hypertabastic model. Due to the hyperbolic functions hypertabastic distribution is rather complicated. I would suggest to the presentation to be less mathematical and to give the reader an understanding why this would be better suited for survival analysis. The formulas don’t offer an understanding of how the distribution looks like. Moreover, the parameters alpha and beta are not defined. The casual reader won’t know what the parameters mean and certainly won’t be able to imagine how the distribution looks like. Moreover, it would be nice to have a proper definition of the mean survival time.

5) Methods section: description of the patients and data need to be improved. Here I would like to see the patient data presented in table format. This is very common (almost the norm) of medical studies.

6) Methods section: the forward selection procedure needs clarifications. Which was the starting variable and what was the inclusion criterion? I don’t see why a correlation of 0.5 would be a problem. Please give the exact specifications of the procedure that ensured that no co-variates had higher pairwise correlation of 0.5. The question is if only one of the pair was included in the model, and which one. How about if a variable had higher correlation than 0.5 with two or more co-variates.

7) Results and discussion: I wish that this section is split in two sections. One clear-cut results sections and one Discussion section. In the present form is a rather heavy mix which makes it very hard to read and take home the message.

8) Results and discussion: Model based on gene expression: I find the discussion and table text for table 3 misleading. The authors claim partly that table 3 gives insights about the accuracy of different models and partly the predictive power. This is not the case as Table 3 presents the log likelihood and AIC values, which are not a measure of accuracy and predictive capability. Likelihood and AIC never give a measure how good a model is, it only can say which model of the considered ones best fits the data. If the best model indicated by AIC have or not a good predictive capability, we cannot know based on likelihood or AIC.

Minor Essential Revisions

1) Results and discussion: Model based on gene expression: Table 2. Perhaps would be better to have only the hazard ratio and confidence intervals. Define the estimated alpha and beta values. Why the authors conducted null hypothesis testing and what was the null hypothesis.

2) Results and discussion: Model based on gene expression: which model did the authors used in the selection? Simply listing it would save you the trouble of going through in length the selection process.

3) Results and discussion: Model based on gene expression: In the next step the authors add to the model a gene expression signature. My question is: why this step is necessary? Why the results weren’t based on the model selection procedure described in the material and methods section.

4) Results and discussion: Time course of survival and hazard: how was it
possible to build a model with ErbB2+ fixed at its media value? A constant does not add anything to a model. Same for Table 10.

5) Core serum response correlation is not defined properly in the text. Given that it is one of the main co-variate and it resurfaces in the figure label 5, it would be nice to have a proper definition.

6) The y-axis of figure 5 (upper right) can be misleading. It would be nice if it could have the same direction as the others. Moreover, what is the point of figure 5. It is difficult to understand how the value at a given correlation of the cumulative survival S(x) can be lower than its derivative S′(x).

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests