Reviewer's report

Title: Assessment of Performance of Survival Prediction Models for Cancer Prognosis

Version: 1 Date: 29 February 2012

Reviewer: Harald Binder

Reviewer's report:

The evaluation of prediction performance of time-to-event models is an important topic, as the difficulties of time structure, including censoring, have frequently led to oversimplifications, e.g., considering only binary endpoints. Therefore, I welcome this manuscript, in principle. However, the present version has some severe issues, which greatly diminish its value and might even be misleading in the conclusions. This needs to be addressed.

- Major Compulsory Revisions

1) You state that to your knowledge "evaluation and comparison among the performance metrics have not been publish". I agree that there is probably no comparison of that specific set of metrics considered in the present manuscript. However, there are numerous other comparisons that consider at least some of the metrics investigated here, and the current manuscript does not cite any of these comparisons. You might, for example, use the recent comparison of Hielscher et al. (2010. Stat Med, 29:818-29) as a starting point for an overview of what has been done already and what could be the specific contribution of the present manuscript to this body of knowledge.

2) The Brier score is missing in the list of performance measures. In contrast to the other metrics, it provides a strictly proper scoring rule, i.e., it takes its minimal value only for the true survival probabilities. Therefore, the Brier score in the course of time, also known under the name "prediction error curve" (Graf et al., 1999, Statistics in Medicine 18, 2529-2545) should be included in the evaluation.

3) I don't understand why you use different model building approaches for the time-to-event and the binary endpoints, namely Cox models vs. SVM/random forest. The closest binary response match for the Cox model would be logistic regression. Furthermore, random forests are available for binary as well as for time-to-event endpoints. As the difference in approaches makes it much harder to compare the results between the different types of endpoints, I would strongly suggest to use the same/similar approaches for the different endpoints.

4) Contrary to the claim in the abstract, I don't really see a reasonably comprehensive comparison of models for time-to-event and models for binary responses. The only empirical result seems to be that for the survival endpoints the clinical model (A) performs better, and for the binary endpoint gene expression (B) performs better. However, given the widely different model fitting
techniques used for the different types of endpoint, it cannot be concluded that this difference ("A better" vs. "B better") is due to the different type of endpoint. For comparison, the analysis in Binder et al. (Biometrical Journal 53, 170-189) keeps the method for model fitting fixed, while changing the endpoint. The results there might serve as an orientation/starting point for the present manuscript with respect to comparing time-to-event and binary endpoints, and should be discussed.

- Minor essential revisions

5) The manuscript distinguishes between a split-sample approach an cross-validation. I would suggest to more generally discuss resampling techniques, which also include bootstrap approaches for evaluating prediction performance.

**Level of interest:** An article of importance in its field

**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