Reviewer's report

Title: External validation of a Cox prognostic model: principles and methods

Version: 1 Date: 4 September 2012

Reviewer: Frank Harrell

Reviewer's report:

Major Compulsory Revisions

This paper has the potential to fill a needed gap, but the authors, champions for effective use of continuous variables, have resorted to dichotomization throughout this paper.

The other general issue I have with the paper is that the authors have included a method related to decision analysis that is not clearly relevant for model validation. Patient utilities play key roles at the bedside when a treatment decision is actually made, but may not be appropriate and certainly not on a par with calibration and discrimination when it comes to model validation. Given the paper is very long, this comment is perhaps especially important.

ABSTRACT: Methods: Line 3 (L3): There may be no two patients with the same risk. Risk groups may not exist or may not be needed, especially if this involves dichotomania. Similar comments for Background L4. The purpose of risk models is not to stratify patients but to estimate risk.

Pg 6 L3: Stepwise variable selection fails even with a very large number of events.

Example, last line: The central limit theorem is not relevant here, and your observation is not in line with many models that I've seen (you don't have to look farther than a model containing only a single binary predictor).

Risk groups: Why are risk groups needed? Why are they relevant?

Example (Pg 7) L4: Creation of heterogeneous prognostic groups is ill-advised.

Pg 8 L2 Line 1: Why bless this common but ill-advised practice?

Next paragraph last line: No, this is not sensible in any way. Risk relationships are not defined by how many patients are like a given patient, i.e., where a patient falls in comparison to other patients.

Pg 9 L3: Why assume basic comparability? Model validations are conditional on covariates, not marginal.

Pg 10 top: Provide a formal test that the slope of the PI in the validation data is 1.0 before proceeding to the more complicated test.
Pg 11 2nd full paragraph L3: strike "ROC index or"

Pg 14 top: Risk groups are a bad idea, and there is no reason that distributions have to be similar between training and test samples.

Before Results: The authors failed to take the comparison of underlying survival curves in training and test samples to its logical conclusion.

Method 1: Relative calibration is more relevant than preserving discrimination in this sense.

Pg 15: Method 5: This is not worth mentioning.

Pg 18 line 11 from bottom: Risk stratification is not what doctors and patients want.

Pg 20: Risk groups are heterogeneous and arbitrary.

Pg 20 bottom: Steyerberg has a paper on estimating the size needed for an external validation.

Figure 6 is difficult to interpret and not fully motivated.

Minor Essential Revisions

Another general criticism, though debatable, is that the baseline survival function for the Cox model is so routinely computed that it is almost part of a standard Cox model fit. The last sentence of ABSTRACT: Background is not strictly true. The 2nd full paragraph on P.3 may also need to be re-worded.

Prognostic index, L2: Such variable selection creates a multitude of problems.

Discretionary Revisions

Pg 4 L1: Is this artificial censoring, which loses information, really needed?

Pg 5 L1: State the overall effective alpha with this strategy.

L4: Why did it need to be eliminated?

age^3 and age^3 x ln(age) does not have face validity

Survival probabilities, last line: It is not unusual to see the underlying S(t) over a fixed grid of times.

Pg 16 before paragraph beginning "We applied the averaging": This is calibration in the large in a sense, and the 3 groups are arbitrary.

Pg 17 middle: A fixed time point is not what physicians or patients want.

Pg 17 line 5 from bottom: Refitting covariates to validation data seems to be a dangerous practice.

Pg 18 middle: The need for replication, when referring to an external sample, is
overrated in some situations.

Pg 21 middle: Cox models deal with relative hazards, not relative risks

Pg 21 Line 5 from bottom: The intention to write another report is not fully relevant here.

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