Reviewer's report

Title: Computerized prediction of intensive care unit discharge after cardiac surgery: development and validation of a Gaussian processes model

Version: 2 Date: 23 August 2011

Reviewer: marieke schuurmans

Reviewer's report:

Thank you for answering to all comments and incorporating most corrections as suggested. Although the paper has improved from its initial submission, there remain questions and limitations.

1. Is the question posed by the authors well defined?

- The matter of a good resource management in the ICU and the possible future added value of clinical decision support tools both clarify the research question sufficiently.

2. Are the methods appropriate and well described?

- Major Compulsory Revision: The proportion of missing data in both the development and the validation cohort is not what is meant with ‘an overview’ of missing data. The 5.18% missing data in the cohort database and the 5.74% in the validation cohort are both mean percentages and do not reveal the dispersion of missing values among the included variables. We would suggest to show the percentages of missing values of the variables with the largest amounts of missing values.

- Major Compulsory Revision: We do agree that replacing missing values would imply changing essential features in the development cohort and relearning of the GP models. The addition in the discussion solves a part of the problem, because it clarifies that there is a limitation in the chosen imputation technique. Unfortunately it does not give insight in which biases are (possibly) introduced. From the statistic point of view, this is what we want to know in order to reason in what way this influences the outcomes of this study and how to interpret the results. We acknowledge this is a difficult problem to solve. As a kind of sensitivity analysis, we suggest to reveal a tip of the iceberg by applying at least one other imputation technique providing new outcomes as a sort of reference. This can be briefly discussed also on the same position in the discussion.

- Major Compulsory Revision: To our judgment the approaches of the peak shaving method and the time-series analysis are clarified. In order to guarantee
the clinical value of the outcomes, subject matter knowledge must always be
leading in all analysis. The addition does not cover this point. We suggest to add
furthermore something like: ‘as doctors do not look at high frequency variation,
but at the overall evolution of a measured signal’.

- Discretionary Revision: To our opinion the reliability diagrams (Figure 2) provide
a lot of information, both essential (observed fraction of positives against the
predicted fraction of positives) and not vital supplementary (the white bars
represent the absolute number of patients, against the predicted probability
divided into 10 bins), which makes it bit confusing. We suggest to remove the
white bars.

- Major Compulsory Revision: The calibration statistics reveal that the GP models
showed to be the only well calibrated models. This is a very important finding.
Because in clinical practice the calibration shows the value of the right prediction
of an individual patient. (As a comparison: discrimination is a measure more on
population level, having a value for managing tasks.) Please point out the clinical
value of this in the discussion! This is a key message (added value) of this
article.

- Minor Essential Revision: Table 3 shows now good information on
discrimination (aROC), accuracy (both Brier statistics) as well as calibration (HS
GOF #2 p-value). We suggest to add also to the table the way it should be
interpreted. Such as higher aROC values, lower Brier Scores, higher Brier
Scaled and higher HS p-values means etc..

3. Are the data sound?

- Minor Essential Revision: In the medical literature, prospective validation of
predictive models in a previously unseen dataset is the most generally accepted
method. Please provide reference(s) to support this statement.

- The way dealing with the deceased patients is now sufficiently explained.

4. Does the manuscript adhere to the relevant standards for reporting and data
deposition?

- A more balanced description now of the loss penalty function, both Brier scores,
Hosmer-Lemeshow statistic and the predicted Gaussian distribution.

5. Are the discussion and conclusions well balanced and adequately supported
by the data?

- Through the removal of the sentence on page 12 this paragraph is more
balanced now.

- The addition on page 13 is satisfactory.

6. Are limitations of the work clearly stated?
- Discretionary Revision: The sentence in the discussion: “Second, the predictions by physicians and nurses might have been biased in a sense that they could have postponed their predictions in the more difficult to predict patients.”, reveals indeed that it is possible that the physicians only in those cases that were clear made a prediction, thus introducing selection bias which can explain the fact that they show the best predictions. The question remains how this influences the outcomes of this study. We suggest to add some kind of answer to this question after this sentence in the discussion.

- The answer to the question concerning the added value using a locally derived predictive model as a basis of an ICU capacity planner above using the EuroSCORE for this purpose, is satisfactory.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?

- These references underpin and support that the EuroSCORE originally developed for the prediction of 30-day mortality is also validated to predict prolonged ICU-stay and is therefore a good reference model for this study.

8. Do the title and abstract accurately convey what has been found?

- Yes

9. Is the writing acceptable?

- Yes

**Level of interest:** An article of outstanding merit and interest in its field

**Quality of written English:** Acceptable

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