Reviewer's report

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

Version: 1 Date: 20 May 2011

Reviewer: marieke schuurmans

Reviewer's report:

In this paper a predictive model for ICU discharge after non-emergency cardiac surgery was developed to predict the probability of ICU discharge the day after surgery (classification task), and to predict the day of ICU discharge as a discrete variable (regression task), by analyzing the first 4 hours of data in the computerized medical record of these patients with Gaussian processes, using a test cohort (n=462) and a separate (previously unseen) validation cohort (n=499).

The researchers soundly attempted to develop models that really can predict the ICU length of stay of an individual patient rather than only predict a prolonged ICU length of stay of an individual patient.

1. Is the question posed by the authors well defined?
   • The background section opens with the statement, "The intensive care unit (ICU) length of stay (LOS) of patients undergoing cardiac surgery may vary considerably.". In this paper it remains unclear why this is a problem.
   • Considering the aim of this paper, to develop a predictive model, it is unclear why the comparisons are made with nurses and physicians estimates. The clinical goal of these models is to support the clinical decision making.

2. Are the methods appropriate and well described?
   • In the method section, Database setup and input data: An overview of the missing data is not provided! For admission data imputing the population mean for missing numerical data yields drawbacks, like regression to the mean. And imputing the value for the parameter that corresponded to a normal healthy condition for missing categorical data, is not free of introducing bias. I would suggest at least to fit a regression model based on the other characteristics to estimate the missing values.
   • For the dynamic data there are some problems in removing the artifacts from the time series. Removing should always be based on subject matter knowledge. It remains unclear if using the peak shaving algorithm (values exceeding 2 times the SD of the time series) are based on subject matter knowledge. This should be explained. Especially because the missing values were also linearly interpolated between two consecutive adjacent known values.
   • First concerning the validation of the models, apart from the ROC-curves, which reveal the discrimination ability of the models, the reliability diagrams (mentioned
in the third paragraph) are missing. In this case of predicting ICU length of stay calibration measures of the models are highly important because these models predict individual length of stays.

• Second concerning the validation of the models, besides a Brier score, which is merely a measure of accuracy, a real calibration statistic (f.i. the U-statistic or the Hosmer-Lemeshow goodness of fit statistic) to support the (not depicted) diagrams is also missing.

• Third concerning the validation of the models, when Brier scores are given to provide insight into the distance the model creates between the predicted probability and the observed probability, one should realize this is also influenced by the median ICU length of stay. And because there is a slight difference between the median ICU length of stay in the test cohort and the validation cohort, I would suggest to provide also the Brier scaled ( (Brier/ mean(model) * 1 - mean (model))/100 ). This scaled version of the Brier score is not depending on population differences of the outcome and gives a more robust comparability of the accuracy of the model in both cohorts.

3. Are the data sound?

• Why is chosen to use a test cohort and a validation cohort is not well described. Another approach was to combine both cohorts and develop the models and validate the models by bootstrap using again the combined cohorts.

• Why is chosen to include death in the ICU as discharge in this study to predict discharge to a normal ward?

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

• There is some imbalance in the comprehensive way the loss penalty function and the Brier score are explained and the very limited way the Gaussian Probability Distribution implemented according to the algorithm is described.

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

• In the discussion is stated that theoretically, including the EuroSCORE parameters in the GP may further improve its predictive power. This statement seems to contradict with the unacceptably high Brier score between the classification model and the EuroSCORE and the statement that it is well known that the EuroSCORE tends to overestimate the operative risk.

• Concerning the last paragraph of the discussion and the conclusion, what is the clinical value of estimating the ICU length of stay as good as the physicians by a computerized model? The physicians are present at the ICU and the physicians decide whether a patient is discharged from the ICU.

6. Are limitations of the work clearly stated?

• Only in 159 of the 499 patients a prediction by physicians was obtained within
six hours after admission to the ICU. In the limitations this is regarded in the light
of the loss of power. Not any consideration is given to the question why in almost
2/3 of the patients the physician was not able to do any prediction within these
six hours. Is it 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.

• Concerning the last paragraph of the discussion, what is the added value using
a locally derived predictive model as a basis of an ICU capacity planner above
using the EuroSCORE for this purpose? Even more models are available and
validated for this purpose.

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

• The EuroSCORE is developed in the early nineties as a model for the prediction
of 30-day mortality. Later the model is also validated to predict prolonged
ICU-stay.

# Ettema RGA, Peelen LM, Schuurmans MJ, Nierich AP, Moons KGM. Prediction
models for prolonged intensive care unit stay after cardiac surgery: systematic
review and validation study. Circulation 2010;122;682-9.
2004;78:1528 –1535.
2003;23:595–598.
# Pitkänen O, Niskanen M, Rehnberg S, Hippeläinen M, Hynynen M.
Intra-institutional prediction of outcome after cardiac surgery: comparison
between a locally derived model and the EuroSCORE. Eur J Cardiothorac Surg.

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

• Yes

9. Is the writing acceptable?

• Yes

Major compulsory revisions

Before decision on publication can be reached it is necessary to the points raised
with regard to the Methods (point 2), with regard to the data (point 3), with regard
to discussion and conclusions (point 5) and with regard to the limitations (point 6)
are addressed.

Minor essential revisions

We trust the author to include the aspects described in point 1, 4, and 7.
Review: Roelof Ettema, Marieke Schuurmans

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