Author's response to reviews

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

Authors:

Geert Meyfroidt (geert.meyfroidt@uzleuven.be)
Fabian Güiza (fabian.guiza@med.kuleuven.be)
Dominiek Cottem (dominiek.cottem@uzleuven.be)
Wilfried De Becker (wilfried.debecker@uzleuven.be)
Kristien Van Loon (kristien.vanloon@biw.kuleuven.be)
Jean-Marie Aerts (jean-marie.aerts@biw.kuleuven.be)
Daniël Berckmans (daniel.berckmans@biw.kuleuven.be)
Jan Ramon (jan.ramon@cs.kuleuven.be)
Maurice Bruynooghe (maurice.bruynooghe@cs.kuleuven.be)
Greet Van den Berghe (greet.vandenberghe@med.kuleuven.be)

Version: 3 Date: 18 October 2011

Author's response to reviews: see over
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:** 29 August 2011

**Reviewer:** Thomas Higgins

**Reviewer's report:**
Thank you for the opportunity to re-review this paper. The authors have successfully addressed issues raised in the prior review. In particular, I was happy to see the percentage of missing data addressed (5 or 6 percent is acceptable) and clarity on how missing data was handled. Imputing normal values is probably the safest course, despite the danger of regression to the mean. With a 5% missing data rate, this should not affect the findings.

I am attaching a .pdf of the Clinical Intensive Care paper, but there is no need to cite it in the article. Unfortunately, that journal ceased publication a few years ago and it is difficult to find electronic copies online.

**Reply:** the authors wish to thank Dr Higgins for his encouraging comments and stimulating insights which have certainly helped us to improve the manuscript.

Minor point; page 4, Methods>Patients: was “thromboembolectomy” or “thromboendarterectomy” intended (versus thrombendarterectomy)

**Reply:** Dr Higgins is right on the misspelling of this word. According to Wictionary, the preferred term is “thrombendarteriectomy”.

I think the paper is much improved with minor editing, the inclusion of the histogram for ICU length of stay, and the addition of primary source references in place of a single review citation. Thank you.

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

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**
I declare that I have no competing interests.
I previously disclosed that I hold 100 shares of Cerner stock (Cerner markets APACHE) but do not feel that Cerner will gain or lose financially as a result of this paper.
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.

Reply: The authors wish to thank Dr Schuurmans for her thorough and stimulating review. We are honoured by the high level of interest score she attributed to our manuscript. In the following section, we have addressed her comments point by point, and we have adapted the manuscript accordingly. Changes in the manuscript, as compared to the previous version, are marked in red.

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.

   Reply: Pre-admission and pre-operative data were the most frequent missing data. Especially spirometry and pulmonary function tests are performed in most but not all patients. We register 6 pulmonary function test parameters in the database (oxygen saturation at room air, vital capacity both in liter and %, one second value both in liter and %, and the Tiffeneau index) of which 31-87% were missing or incomplete. The following sentence was added to the methods section on page 6: “Variables with the largest amount of missing data were mainly related to pre-admission assessment of the patients, such as the pre-operative spirometry and pulmonary function tests (31-87% missing), pre-admission heart rate (35% missing), pre-admission blood pressure (27% missing). Intra-operative blood loss was not entered in the database in 50% of the patients.”

   - 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.

Reply: upon request by the reviewer, this analysis has been done. We have replaced the missing values in the numerical admission data category (because that category had the largest amount of missings) by the values of that parameter that corresponded to a normal, healthy condition, instead of using the population mean. The model for the admission data was relearned, and we observed no significant difference in performance (aROC 0.73 when we replaced by mean values, went to 0.72 when using normal values) . We have added the following sentence to the discussion session on p 12: “Nevertheless, when a different imputation method was used, such as replacing missings by values corresponding to a normal healthy condition, the results did not change significantly (data not shown).”

- 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’.

Reply: we thank reviewer 2 for this excellent suggestion, and have added the following clause: “very similar to the way doctors look at continuous parameters.”

- 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.

Reply: upon request by the reviewer, figure 2 was adapted, and the white bars were removed. The legend of figure 2 was adapted accordingly.

- 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.

Reply: upon suggestion by reviewer 2, we have adapted this paragraph in the discussion section on p 12: “At the population level, calibration and accuracy of the second day discharge predictions was good, with aROC well above and Brier score well below the predefined thresholds. At the individual level, the GP models showed to be the only well calibrated models. This is extremely important in clinical practice, when using the models for patient counseling.”

- 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..

Reply: The interpretation of these measures is explained in the methodology section. As requested, a reference to where this information can be found in the manuscript.
was added to the legends of table 3, as well as to table 4: “Legend: The interpretation of the different validation measures can be found in the methodology section.”

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.

Reply: External validation of models is important with regards to generalizability. As requested, a new reference 20 was added: “Steyerberg EW: Validation of prediction models. In Clinical prediction models: a practical approach to development, validation and updating. New York, Springer; 2009: 299-311. “
- 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.

Reply: the following clause was added: “whereas the GP models have always delivered a prediction within the allotted time, regardless of the uncertainty. “
- 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.