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 (mauricebruynooghe.cs.kuleuven.be)
Greet Van den Berghe (greet.vandenberghe@med.kuleuven.be)

Version: 2 Date: 5 August 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: 1 Date: 21 April 2011

Reviewer: Thomas Higgins

Reviewer's report:

This study developed a predictive model for ICU discharge after cardiac surgery via a Gaussian processes model. Most prior work has used logistic regression. Limitations include a single center study, and a relatively small n of 461 in the development set. The authors acknowledge this limitation in potential generalizability (page 12). The ROC is moderate at 0.758 (realizing that ROC for LOS is generally lower than for mortality predictions). Median ICU LOS is fairly long by US standards, which raises the question whether patients might have left earlier. It would be helpful to report the standard deviation of LOS, or better yet display a histogram.

The authors thank Dr Higgins for his thorough review, and his valuable and much appreciated comments. We have addressed his concerns point by point in the following section, and have highlighted the requested changes in the manuscript where appropriate.

Major compulsory changes:

Add a histogram of ICU length of stay, at least by days if not hours

Reply: as requested, a new figure 3 was added to the manuscript, showing the actual distribution of ICU length of stay by days, as well as the distribution of the predictions by the GP models and the EuroSCORE model.

Page 5, discussion of missing numerical data: please cite the % of missing elements in the results section

Reply: the % of missing data in the development and validation cohort was added to the methods section on p 6: “Overall, in the development cohort database 5.18% of data were missing; in the validation cohort the proportion of missing data was 5.74%.”

I was unsure of what was meant by the statement on page 11 that “Validated applications to examine these databases are currently lacking”.. Certainly kappa analysis can be employed, as well as other measures of intra-observer variability.

Reply: we agree with reviewer 1 that this sentence might have been confusing. Therefore, it was removed.

Page 11 - I am aware of at least two studies that have looked specifically at ICU length of stay in ICU patients. (Higgins TL et al, Predicting prolonged intensive care LOS following CABG. Clinical Intensive Care 1999; 10:175-182; Becker et al J Cardiovasc Surg 1995; 36:1; plus others referenced in the Kramer review). To my recollection, these studies also reported aROC metrics on performance, which may have been better than the EuroSCORE. Consider quoting these other sources. As you may know, APACHE has different equations for general ICU use versus customized equations for cardiac surgery, to accomodate the physiologic differences and differing independent variables of interest in the cardiac surgical population.

Reply: as requested, a reference to the paper by Becker et al was added as reference # 2
Minor Essential"
I would move the discussion of machine learning techniques from background to
discussion.
Reply: the authors sincerely apologize but it is not clear to us which section of the
background should be removed to the discussion? In the background part of the
manuscript, only 2 short sentences refer to machine learning techniques ("Machine
learning techniques can be used to analyze such a database in an automatic way [7].
In retrospective database analysis in a surgical ICU population, they have been
shown to predict ICU LOS [8].").
Either a reference or statistical review should justify the statement of 0.25 as an
acceptable cut-off for the Brier score.
Reply: the book by Steyerberg et al, justifies this statement. It is now referenced as
[25].
On page 8, I did not see HIS (health information system) previously spelled out -
if this is the case, the abbreviation needs to be spelled out at first mention.
Reply: HIS (Hospital Information System) was first spelled out in the methods
section, Database setup and admission data, Admission data, on page 5.
Student t-test should be capitalized (page 8)
Reply: was capitalized, as requested
The multiple statistics on page 10 would be easier to absorb in tabular form -
could refer to the table here.
Reply: in order to comply with this comment, the sentence “The results are
summarized in table 4” was put up front on a separate line, in order to emphasize
that the results can be viewed in tabular form as well as in the text.
Discretionary:
Reference 1 is a review of other studies - a very good review, but consider
adding two or three other primary references (see comments re: page 11,
above). Reference 2 is also a review, and I would consider separately
referencing the scores cited (EuroSCORE, STS, Parsonnet, etc.) rather than the
review, especially since readers of this journal may be less familiar with the
original work, and may need access to the detailed statistical descriptions in the
original articles.
Reply: as requested, these references were added: references # 3,4,5,6,7,8.
On Page 6, the paragraph on dynamic data is very long, and the last two
sentences may be too much information.
Reply: upon suggestion by reviewer 1, the sentence “A high ApEn reflects an
irregular time series, a low ApEn refers to a time series with many repetitive
patterns.” was removed because indeed, it was redundant.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: Yes, but I do not feel adequately qualified to assess the
statistics.
Declaration of competing interests:
I hold 100 shares of stock in Cerner Corporation, which licenses the APACHE
system, which is referred to in the paper. I do not consider this a competing
interest, but mention it for completeness.
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.

Reply: the authors thank Dr Schuurmans and Dr Ettema for their extensive and valuable comments. We responded to the remarks in the section below, and have adapted the manuscript accordingly. The adaptations in the manuscript are highlighted in red.

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.
   Reply: the following sentence was added to the introduction, p3: "This is problematic, not only for counseling patients or their relatives on the expected ICU LOS, but also with regards to bed and resource management in the ICU."
   • 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.
   Reply: we agree with the reviewers that the clinical goal of these models is to support decision making, namely planning of future cardiac surgery cases. The current "model" to predict bed availability, or ICU discharge, is the clinician. We wanted to compare the GP model with the current way of predicting bed availability. A clinician would probably question the value of any model he or she can easily outperform. In order to clarify this, we added the clause "....in order to have an indication whether in the future they could have a possible future added value as clinical decision support tools."

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!
   Reply: also upon recommendation by reviewer 1, we added the proportion of missing values in development and validation cohort to the methods section on page 6.
   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.
   Reply: we acknowledge that any method of inputting missing data has its drawbacks. A new method for replacing missings would imply changing essential features in the development cohort and relearning of the GP models. We agree with the reviewers
that we should have mentioned these drawbacks as a limitation in the discussion section. In order to comply with the concerns by the reviewers, we have added the following paragraph to the discussion on page 11: “Several methods exist for inputting missing data, each of them has specific drawbacks. Replacing missing numerical data by the population mean for that value, as was done in the present study, might lead to regression towards the mean. For categorical data, the value that corresponds to a normal healthy condition was used to replace a missing value, which might have introduced some bias."

• 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.

  Reply: in order to address this concern by the reviewers, we will explain how peak shaving was done. In a first step, the trend of each signal was determined with a low pass filter. In the second step, the upper and lower boundaries of the 2SD band centered on the calculated trend were determined, and in a third step, peaks outside this valid region were eliminated. By visual inspection of signals by expert clinicians, we found that these criteria eliminated the spikes and noise due to patient movement, flushing of the lines or disconnection of the transducers.

We acknowledge that we might have failed to provide sufficient detail in the methodology section regarding the artifact removal and data inputting. We understand that the reviewers have some concerns regarding the possible loss of valuable high frequency components in the time series that might have been predictive. As we will explain below, because of the way the time series were summarized, we are not concerned with these high frequency components.

  The time-series analysis, as was done in the present study, has functioned as a low-pass filter: before analysis, the signal was de-noised by removing high frequency components. This was done at 3 levels:

  1. The way data are stored in the MetaVision PDMS database is a low pass filter in itself, because data are stored at 1 minute intervals. This way, high frequency analysis such as beat to beat heart rate analysis is not possible. This is an inherent property of the MetaVision PDMS database which cannot be changed.

  2. We acknowledge that linear interpolation and peak shaving will also eliminate some high frequency components of the time series.

  3. We have applied a third low pass filter on top of that by averaging the signals: in the category "Physiological data", we used mean and variance of the monitoring data. In the category dynamic data, values were averaged in 6 blocks of 40 minutes each. We have assumed that high frequency variation is not predictive for length of stay, but only the trends of the signal. We have based this assumption on clinical experience, as doctors do not look at high frequency variation, but at the overall evolution of a measured signal. If predictive information would be present in high frequency components, we are obviously missing it. Whether adding high frequency components to the predictive models would increase their predictive performance is a very interesting question, which should be addressed in a future study.
We have added the following paragraph to the discussion section on p11: “Time series analysis in the present study was done after applying several low pass filters on the signal that have removed all high frequency components. We have done this under the assumption that the trend of a time series is more predictive for outcome than high frequency variability.”

• 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.

Reply: the reliability diagrams were added (Figure 2). We refer to this figure in the last line of page 9 (results section). (For the interest of the reviewers: in the original first draft of the manuscript, we already included these diagrams, but we had to remove them upon request from one of the first reviewers for Critical Care (which is also a BMC journal).

• 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.

Reply: upon request by the reviewers, we have performed the U-statistic, and have added it to the results section of the manuscript on pages 9 and 10, as well as to table 3, and the abstract. The GP models turned out to be the only well calibrated models. The discussion was adapted according to these findings: (p 11): “GP models were the best calibrated models, whereas the EuroSCORE, ICU nurses and physician’s predictions showed overfitting.”

• 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.

Reply: we have calculated the Brier Score scaled, and added it to the results (pages 9 ), as well as to table 3.

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.

Reply: the following sentence was added to the methodology section (p4): “In the medical literature, prospective validation of predictive models in a previously unseen dataset is the most generally accepted method. Therefore, …..”

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

Reply: there are 3 ways of dealing with patients that have died with regards to length of stay: first, consider death as a way of discharge; second, to censor for patients who died by attributing them a hypothetical LOS beyond the longest staying patient discharged alive, or third by omitting these patients from the analysis. In outcome studies, the 2 latter methods are to be preferred. Because the number of non-
survivors was extremely low, and because we wanted to predict bed availability, we have considered death also as a way a bed becomes available.

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.
   
   Reply: the section on the Brier score was shortened (but a mention of the Brier Score Scaled and Hosmer-Lemeshow statistics was added). The explanation of the predicted Gaussian distribution was also lengthened.

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.
   
   Reply: This sentence on p12 was removed.

   • 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 patients is discharged from the ICU.
   
   Reply: The following sentence was added to the discussion on p 13: “Although the discriminative power of the GP models is not significantly higher than the physicians, they can be of added value because they will deliver their predictions in a more reliable and consistent way (not postponing the predictions in the most difficult cases as the physicians probably did in the present study).”

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.
   
   Reply: Indeed, we tend to believe this is the case. As the reviewers have probably remarked, the GP models have a tendency to perform better in this subpopulation. We did already mention this in the limitations section on page 13 “. 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.”

   • 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.
   
   Reply: the advantage of a general risk score is the generalizability. The disadvantage of using the risk score for prediction outside a setting that is very similar to the setting where the score was developed, is that, without recalibration of some or all of the coefficients used by the model, the predictive power will be significantly lower than the original model. This has been shown for several ICU and cardiac surgery risk
models. The advantage of the locally developed model over the EuroSCORE is predictive performance, as we have demonstrated in the present study.

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.
   Reply: upon suggestion by reviewers 2, these references were added to the “background” section as ref #10, 11, 12, 15

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