Author’s response to reviews

Title: Development and validation of a novel prediction model to identify patients in need of specialized trauma care during field triage: design and rationale of the GOAT study

Authors:

Rogier van der Sluijs (r.vandersluijs@umcutrecht.nl)

Thomas Debray (T.Debray@umcutrecht.nl)

Martijn Poeze (m.poeze@mumc.nl)

Loek Leenen (l.p.h.leenen@umcutrecht.nl)

Mark van Heijl (markvanheijl@hotmail.com)

Version: 1 Date: 22 Mar 2019

Author’s response to reviews:

Dear Editor-in-Chief,

We would like to thank both reviewers for their extensive reviews of the manuscript and the suggested revisions. We adjusted the manuscript to incorporate their suggestions and elaborated our reasoning per comment in the section below.

Reviewer #1:

1. First, the authors might want to address the imbalance of the binary outcome classes by over- or undersampling or even more sophisticated methods for imbalanced classification problems. These approaches can also be part of the tuning process to optimize the sampling of classes.

Thank you very much for the advise. We considered class imbalance in an earlier stage and also (vaguely) incorporated it into the original manuscript (Table 2). We will evaluate two scenarios, one with class weights that are inversely proportional to the class frequency in the input data, and one without class weighting based on the a review of balancing techniques for this family of prediction models.[1] A section to clarify this was added to the manuscript on lines 302-304. We also chose a measure of pure discrimination that is free of prevalence (c-index) and thresholding, and thus is not likely to be impacted by class imbalance. We do suspect that all suggested balancing techniques will lead to distorted calibration, and might not be the better option without recalibration.
2. Second, the main idea of the study is to replace a "simple" model by a more sophisticated one. Still the authors chose to investigate only a single alternative, which is the gradient boosting decision tree. This is a severe limitation. Though it could be discussed as such in the "Limitations" section, I strongly recommend performing a benchmark study of several approaches, for example including Lasso, Random Forests, Neural Networks, Support Vector Machines etc. Otherwise the authors risk ending up with a model which is too simple again. There are several software implementations, for example in R, that provide all the tools one needs to perform such a benchmark study.

We agree with the reviewer that there might be a prediction model, or even an ensemble of different families of prediction models that would lead to a (slightly) higher discriminative value. However, we chose to pre-define the prediction model because of several reasons. First, we suspect that the gradient boosting decision tree will be one of the top results in terms of discriminative value based on the literature and test runs on one of our previous datasets.[1, 2, 3] Second, our goal is to replace the reference model with the newly developed model in practice, a situation in which input data is often incomplete. Gradient boosting machines learn how to handle missing values during the training process, while simple imputation strategies do not and might generate worse results in practice. Third, a considerable part of the model development strategy is focused on generalisability and our implementation of the gradient boosting machine offers more options to prevent overfitting, in comparison to some of the other suggested models. A section to clarify our reasoning was added in the manuscript on line 351-369.

3. Line 83: "All hospitals …" Is it a reasonable assumption that all hospitals will participate in the study and that all outcomes will be available?

Yes. All hospitals with trauma receiving emergency departments are obliged by law to collect data for the National Trauma Registry (in Dutch, Landelijke Traumaregistratie) and our proposal to retrieve this data was accepted earlier.

4. Line 85: "Patients that die during transportation will be excluded". To my understanding it is possible that the death of such patients is related to a wrong decision on where to go. The authors should discuss whether use of a combined endpoint "death or ISS >= 16" is more meaningful.

We agree with the reviewer that a combined endpoint would likely be more meaningful, but originally chose for the sole exploration of ISS >= 16, since this is the suggested way to evaluate trauma systems (and triage) nationally and internationally. Based on this review and the commentary [4] on our previously published model [5], we decided to add a second reference standard which is based on resource-use. We adjusted several sections in the text to reflect these changes: lines 12-15, 99, 115, 148-154 and 347-350.

5. Line 131: Please do also state that variable selection will not be related to the prediction performance, if not performed within the training data.
Added this on lines 164-165.

6. Line 132: How will the additional features be "engineered''?

This was originally described on lines 149-151. An additional section to clarify this was added on line 166-168. Basically, we will convert variables such as date to more relevant predictors, such as season of the year, day of the week and daytime or night.

7. Table 1: I don't see the importance of the "anticipated functional form". Why is this information needed?

We agree with the author that is has no additional value. This column was therefore removed from the table.

8. Line 197-198: It is not clear which "studies" it is referred to.

Clarified this on lines 316, 318 and 319.

Reviewer #2: This is a well-written paper describing the design of the 'GOAT' study. I have only minor comments.

1. Line 60. Please provide some details regarding the 'relatively small sample size'.

Updated this line (72-73) to reflect the actual number of patients included in the previous study.

Line 76. Update this sentence.

This sentence was updated.

Line 147. Consider discussing the (implicit) missing data assumptions made by the machine learning algorithm.

We added a section on lines 214-216 that briefly outlines the assumptions of the machine learning algorithm.

Lines 163-165. Please provide some references for this statement.

References were added here.

Lines 165-175. This section does not provide much insight into how boosting works. In addition, it contains a lot of jargon. Consider expanding this section (with suitable references).
We agree with the reviewer that this section could be a lot clearer. We replaced the section with a new text with a more basic explanation of the boosting principle, suitable references and without the use of jargon. The newly added section ranges from line 248 to 253.

Lines 181-184. This section could be clearer. Consider expanding this section.

Expanded this section on lines 293-296.

Lines 184-186. Please provide a reference for this algorithm.

References were added here.

Line 191. Please provide a reference for the decision curve analysis.

A reference was added on line 305.

Line 235-237. I did not understand this sentence. Please clarify.

Simplified this sentence and removed the technical details.

References:


