Author’s response to reviews

Title: Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records

Authors:

Merijn Beeksma (m.t.beeksma@let.ru.nl)

Suzan Verberne (s.verberne@liacs.leidenuniv.nl)

Antal van den Bosch (antal.van.den.bosch@meertens.knaw.nl)

Enny Das (h.das@let.ru.nl)

Iris Hendrickx (iris@i-hx.nl)

Stef Groenewoud (stef.groenewoud@radboudumc.nl)

Version: 2 Date: 26 Sep 2018

Author’s response to reviews:

Dear Dr. Brian Wells,

Thank you for your continued interest in our research and the opportunity to revise our paper, titled ‘Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records’, that we hereby submit. As you may recall, the reviewers were positive about the paper and had some comments and questions. We thank the reviewers for their insightful comments, which we believe have helped us to improve the paper.

Below, we address each review separately, and respond to the comments in a point-wise fashion. We quoted the reviewers’ comments, leaving out only the reviewers’ summaries of the paper and one repeated comment in review 1.

Each of the reviewers addressed decisions we made in the design of the study, and interpretations of the results. We have processed all comments as explained below, and additionally clarified our choice to aggregate data over one-month periods in the paper (in the second paragraph of section ‘Creating input data for the model’).
We hope our paper is now suitable for publication in BMC Medical Informatics and Decision Making. If anything remains unclear, we are happy to answer questions or consider further revisions. We look forward to your receiving your response.

Yours sincerely,

on behalf of all authors,

Merijn Beeksma

REVIEWER 1 (Shahryar Eivazzadeh)

REVIEWER: “Improving ACP and palliative care can be an output to using your models and methods but as far as I can see your study is not limited to this context and the results can have other applications (like better procurement of care facilities). But, the wording in the abstract and introduction implies that the study is specific to this application, which is not.”

RESPONSE: We rephrased and moved the order of several sentences to broaden the focus and mentioning the more generic applicability, before introducing ACP. We made the following changes:

• instead of directly introducing ACP, the background section of the abstract (p. 2) and first paragraph of the introduction (p. 3) now start with: “Life expectancy is one of the most important factors in end-of-life decision making. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning.”;

• in the first paragraph of the conclusion (p. 28) section, we additionally provide an example of a useful application: “The potential use of automatic prognostication is not limited to health care in practice, but could also be useful in other clinical applications such, such as clinical trials. In clinical trials, outcomes often depend on prognostic factors. Automatic processing of medical
records would enable quick and systematic stratification of patients based on their prognoses, which could be used to further reduce biases [56].”

REVIEWER: “Your model was fed with data of deceased patients (100% in 5 years), but the human agents (doctors) were not sure about death of patients within the next 5 years period (92% for the next 6 month in hospice). I wonder if this difference should be considered when comparing the results (it might slightly favor the machine learning approach)?”

RESPONSE: Maximum life expectancy is limited to roughly four years in our study, and is technically not limited in the comparison study. However, since the comparison study took place in a hospice setting, life expectancy was in fact on average much shorter in the comparison study than in our study. We know from research that it is increasingly difficult to make accurate prognoses about life expectancy as the moment of death lays further away in the future, therefore we do not think that the machine learning approach is favored in the comparison, but rather believe that the task presented to our model was relatively hard. We added this line of reasoning to:

• the final paragraph of ‘Comparison to human performance’ in the ‘Discussion’ section (p. 26).

REVIEWER: “The median survival for your reference of correct prediction [...] was 24 days. As the criterion of correct prediction (33%) gets more generous for longer survivals, then I think comparing accuracy of predictions to the reference is acceptable only if the median survivals are comparable. I could not find median survival days in the manuscript, please indicate that, and if it is very different from your reference you may want to clarify why still you compare the results while the survival median in your study is longer (if longer).”

RESPONSE: Indeed, the medians of survival differ between the studies. However, given that predicting life expectancy is increasingly hard when the actual moment of death lays further away in the future (see previous response), we think that the relative error margin we adopted from the comparison study corrects the differences between average life expectancy in the two studies at least partially. We edited the text in the following ways to explain this:

• in the second half of section ‘Predicting life expectancy with an LSTM’ (p. 8), we explain that there are no baseline systems or benchmark data available, and that a perfect comparison is therefore not possible at this moment;
• we added the medians of survival for both studies the third paragraph of section ‘Evaluation protocol’ (p. 18): “In the hospice setting, 92% of the patients lived for maximally six months after admission, and the median of survival was 24 days. In our study, the maximal life expectancy was roughly four years, or fifty months. The chances of dying were evenly distributed over these months as a result of the sliding window approach, thus the median of survival was 25 months.”;

• in the third paragraph (p. 18), we emphasized the differences between the studies, and in the fourth paragraph (p. 18), we explain the similarities between the studies, and our motivation for including the comparison;

• we added section ‘Comparison to human performance’ to the discussion (p. 25), in which we reflect once more on the comparison;

• in this section, we also added references to research reporting on the relation between prediction accuracy and time to death (p. 26): “However, as prognostic accuracy tends to be inversely related to a longer life expectancy [16, 54, 55], we assume that the task we formulated was relatively hard compared to the task presented to the doctors: because life expectancy was uniformly distributed over 1-50 months in our research, the model had to make predictions about the near future (one month into the future) as well as the far future (fifty months into the future).”;

• in the final paragraph of the conclusion, we explain that we will address a direct comparison with human prognostic accuracy in future work (p. 30).

REVIEWER: “P7L9-13: please provide reference for (probably #22 ?)”

RESPONSE: We added this reference.

REVIEWER: “P16L44: do you mean "forward stepwise " feature selection approach? As you did not tested all combinations, then you may want to motivate why you preferred this approach.”

RESPONSE: We indeed meant forward stepwise feature selection. We added an explanation about why testing all possible combinations was not feasible in:

• the final paragraph of section ‘Selecting features for the structured EMR data’ (p. 15): “Testing all selections of features would have made the grid search infeasible, therefore we
determined redundancy on the level of feature categories. We used a forward stepwise feature selection approach: [...]”.

REVIEWER: “P27L9-13: this is an important extension to your work, you may want to elaborate more on the requirement of this "more data".”

RESPONSE: We added to the conclusion that we are currently collecting more data, and explain what we want to do with this data in future work, in:

- the final paragraph of section ‘Conclusion’ (p. 29-30): “In order to replicate and extend this research, we are currently expanding the dataset substantially, by collecting additional data of both deceased and active patients. This will allow us to zoom in on specific illness trajectories, and to rephrase the task in such a way that it will match clinical settings more closely, for example by aiming to make predictions about patients while they are still active.”

REVIEWER 2 (Christine Fennema-Notestine)

REVIEWER: “The study is of interest, and the findings intriguing in an exploratory sense. As the authors note, however, the amount of data "is not considered to be a lot of data for training neural networks." Despite that note, there are a number of sub-explorations of conditions. The findings must be taken as exploratory and an emphasis placed on the need to replication and extension of this work in the manuscript.”

RESPONSE: We emphasized the exploratory nature of the research and provide a number of ideas for extending the research in the following section:

- the final paragraph of section ‘Conclusion’ (p. 29-30): “Nevertheless, this research should be considered to be exploratory.”

REVIEWER: “In addition, the background information prior to methods presentation could be reduced and streamlined.”
RESPONSE: We agree that the background section was quite extensive, especially with regard to the background information on ACP, because although we place our research in the context of ACP, this is not the main subject of the paper. We removed a few paragraphs and rewrote parts of the background section, to streamline it (without removing references to relevant literature):

- The background section up to subsection ‘Automatically processing clinical data’ (p. 3-4).

REVIEWER 3 (Mohammad Reza Daliri)

REVIEWER: “As mentioned by the authors "Life expectancy is one of the most important factors in determining the right moment to start Advance Care Planning" and "Physicians however tend to overestimate life-expectancy" but in my opinion the authors results is not enough and much better than the Physicians (based on the results reported in the tables in the results section).”

RESPONSE: Because good prognostication has the potential to contribute significantly to end-of-life decision making, we propose that any increase in prognostic accuracy is worth the effort. Additionally, even if no improvements would have been made, a system for automatic prognostication is useful because large amounts of data can be analyzed at once, without the need of (human) medical expertise. However, we did not intend to suggest that we have solved the problem by increasing the accuracy by 9% compared to human prognostic accuracy. We added:

- the fourth paragraph of section ‘Conclusion’ (p. 29): “Even though the model’s performance is far from perfect, we consider this work to be among the first steps in a line of research that has much potential for clinical applications, for several reasons: good prognostication has the potential to contribute significantly to end-of-life decision making, therefore we believe that any increase in prognostic accuracy is worth pursuing. Additionally, human prognostication is costly, time-consuming, requires medical expertise, and is a subjective task. Without compromising prediction accuracy, the model is able to make predictions quickly, automatically and systematically, while it does not depend on human medical expertise. Even though the model reaches only 29% accuracy, we consider 9% point improvement to be promising, considering that the model is trained on a relatively small data sample.”

REVIEWER: “In addition the logic behind using different part of the model is not clear.”
RESPONSE: We assume the reviewer refers to the text processing steps we applied to the clinical texts, and we added a few examples to give an indication of the adjustments we made to the texts, which can be found in:

- the final paragraph of subsection ‘Unstructured data’ in the ‘Method’ section (p. 11).

REVIEWER: “There are many approaches in Machine learning which can be used for prediction and why the authors have used long short-term memory recurrent neural network? Maybe other statistical learning approaches (like different Bayesian approaches) might work better in this context.”

RESPONSE: We agree that a comparison with other machine learning algorithms would be interesting to investigate in future research. However, an LSTM model was especially suitable for our task, because it is able to process sequential information and, in contrast to other machine learning algorithms, is optimized to learn to keep long-term dependencies in memory, which is a feature that is relatively unique of LSTMs in comparison with other sequence processors (e.g. Conditional Random Fields or simple Recurrent Neural Nets). We explicated this motivation in the following section:

- First paragraph of ‘Predicting life expectancy with an LSTM’ (p. 8).

REVIEWER: “The evaluation has been done on a small sample of the data (1234) which might not be easily generalized. In addition the way of evaluation and selecting the test data might not be a good one. The three-way of split the data is proper (train-validation-test) but the this should be repeated many times over the whole data. It seems that the authors have selected the 10% of the test data once (the 10% most recent patients from health care facilities). This should be shuffled over the whole database similar to what has been done for the train and validation set. (the average and standard deviation of the results then can be reported).”

RESPONSE: The alternative approach suggested by the reviewer is indeed proper and could be followed. However, we would like to argue that our approach constitutes a proper evaluation protocol as well. We purposely set apart 10% of the data in order to mimic a real-world case, in which a trained model would be applied to entirely new data - data which at no point has entered the cross-validation cycle (which we performed on the remaining 90%). To emphasize why we set 10% of the most recent patients apart, we added an explanation to:

- the first paragraph of subsection ‘Train-validation-test split’ (p. 11).