Reviewer’s report

Title: Importance of Medical Data Preparation in Predictive Modeling and Risk Factor Discovery for the Frailty Syndrome

Version: 1 Date: 14 May 2018

Reviewer: Klaus-Hendrik Wolf

Reviewer's report:

The shortening of the paper has resulted in a better readable article.

The presented work pursues three goals. (a) The use of data from EHR for medical data analysis, (b) the application of the CRISP-DM method to medical data, and (c) an exploratory analysis of the study data with regard to frailty predictors and a machine learned model.

First my main concern with the paper.

In the objectives electronic health records (EHRs) are not mentioned.

Usually, if it is in the title it should be in the objectives.

But...the data sources used in the article are no EHRs!

Electronic health records in the field of medical informatics are defined as the digital means to hold information of patients from their medical encounters.

The data used in the paper stem from an epidemiologic study.

That means that the data and the procedure of their capturing are highly standardized.

Given the data's source the structure and completeness varies to a large extend from typical EHR data.

So to extrapolate from the findings with this study-data to EHRs in general is at least questionable.

Therefore, the authors should think about whether substracting EHR (goal (a)) from the paper could further streamline the research question and improve the clarity of the paper.

The frailty status of a person is assessed by the Fried's frailty score.

In the further analysis the classes 'pre-frail' and 'frail' are combined.
The authors write on page 3 citing Fried concerning the definition of frailty: "A standardized definition has not yet been established."

The frailty syndrome is quite a diffuse concept making it a blurry target for any statistical analysis. Of course this does not get better if you combine 'pre-frail' and 'frail' into one class.

To find good indicators focusing on the 'frail' vs. 'non-frail' could result in a clearer outcome.

Overall frailty is not the ideal target value to choose, if you want to test a method like CRISP-DM.

Applying data mining to the presented data set the authors correctly exclude the main known predictors for frailty from the mining process, as the diagnosis of frailty was based on these values.

The problem with this approach is, that you are not able to learn a model that is useful in the real world, as the prediction will get better, if you use the defining variables.

So the added value is that the process might come up with further indicators for frailty.

This should be clearly stated in the goals.

Regarding testing CRISP-DM it is not clear, what the paper adds to the existing knowledge.

The authors give an overview of the development of the CRISP-DM methods since its start in 1996.

They even cite the CRISP-MED-DM model by Niakasu, but it is not explained to which extend it is implemented in the study.

Currently the paper lacks a short informative description of the CRISP-DM method and the steps defined by this method.

The main conclusion here seems to be that you can apply CRISP-DM to medical data?

Regarding the imputation, it is at least questionable to first impute using correlations between variables and then search for correlations.

The imputation will boost existing correlations and you have to at least compensate for that and describe how you did that in the paper.

A section on limitations would be useful.
To summarize the paper currently reports towards three goals.

I would strongly recommend to focus on one or two and formulate clear research questions.

Please consider to write a section on limitations and clearly state what the paper adds to the state of the art.

**Are the methods appropriate and well described?**
If not, please specify what is required in your comments to the authors.

No

**Does the work include the necessary controls?**
If not, please specify which controls are required in your comments to the authors.

Unable to assess

**Are the conclusions drawn adequately supported by the data shown?**
If not, please explain in your comments to the authors.

No

**Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?**
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I recommend additional statistical review

**Quality of written English**
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

**Declaration of competing interests**
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal