Reviewer’s report

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

Version: 0 Date: 31 Oct 2017

Reviewer: Marcel A L M Van Assen

Reviewer's report:

Am used to reading articles of about 20 pages, with a certain structure (intro, methods, results, conclusion/discussion) where each section does not overlap with others. The paper I reviewed does not resemble articles that I use to read and review, which explains many of my comments below.

This paper is more of a book than an article. I believe the paper can be shortened very, very much; I provide some suggestions for shortening the paper below. On the other hand, sometimes I believe some vital or interesting information is missing; again I provide below which information. Finally, the structure of the paper is not always obvious to me; for instance it seems that the method section includes some introduction material and some results. So, the structure may be streamlined as well.

My comments in detail:

Abstract: Delete 2nd sentence (always important topic). Fourth sentence: rephrase or delete, because postponing bad outcomes may even increase health costs if people live even longer. Background section is (too) long. Methods section is unclear: what parameters (list kind of variables)? CRISP-DM? data understanding and pre-processing phase? EHR? Results: first part is repetition, second part - what is the actual contribution (as I do not see anything particularly new)?

Writing: Please check English and grammar before submitting again. Please do not write paragraphs of more than one page with multiple messages. For instance, the first paragraph of the Background is much too long and is better divided into multiple paragraphs, each with its own message.

P3, 23: Make clear the design, i.e., list the variables that were assessed and when. This is vital information.
Objectives: "This model could …" does not belong to the objective.

Related work 1: A paper is not a book. This section is much too long (almost 20 pages). Readers do not need to know from your paper what each individual other paper did. It is enough to sketch what clusters of papers did. Vital is to know WHAT clusters of papers did (how many observations, which variables [note that even though your section is very long, this vital information is often missing], time frame), and what can we conclude from all these papers. All in all, one or two pages background should be sufficient for the whole section until the frailty subsection. As this paper is about frailty, that section may be a bit longer.

Related work 2: I do not know what c-statistics are (P10,4). How do you define accuracy (P12, 15). "16,000 cases and 283 features": please always list the kind of variables used, because this line is not informative. This holds for the whole paper (e.g. P18,17).

Frailty in Background: (i) 377 features □ WHAT kind of features/variables? This is not helpful. (ii) I do not understand many bullets of the list on page 20 (1st, 4th, 7th). Please indicate the sign of each relation. (iii) Why so much info on Swindell et al relative to Baylis et al as the latter paper was so important (according to Baylis, at least)?

Common analysis techniques: (i) what are CERs and mProbes?; (ii) please limit yourself in literature and text, preferable to that what is directly relevant to frailty and the purpose of your study. This section seems to discuss some studies that do not seem directly relevant for the present paper.

Method description: (i) life cycle refers to? (P22,26). What is the relevance of 'business understanding' and 'understanding frailty problem' in the methods section? The 'understanding frailty problem' seems to be more suitably located in the introduction section (relevance of research). (ii) Why is this section called "data understanding", and not "data" or "data collection"? (iv) what are "anthropometric data"? (v) definition of variables: is there a list, possibly in supplementary materials, which explains in detail how each of your features is assessed? (v) the quality of some of the data seems questionable (e.g. P28,28-30). Why? This, and its implications, are important topics for the discussion. (vi) What is exactly the use of the PCA here? If you have a different number of variables per "group", the first two PCAs likely represent the "group" with the most variables per group. So, I clearly see the added value of the
PCA to group variables, but I do not see (yet) the added value of showing the results of the first two components. Note also that the third component explains almost as much as the second, so there indeed does not seem to be any reason not to show the results of the third component. (v) For the k-means clustering, were the variables standardized before the analysis or not? What is the interpretation of clusters 1 and 4, as they do not seem to differ on the relevant variables? Are all figures needed here, e.g. the figure with distribution of sexes seems to be superfluous. And what is the information that we gained from the cluster-analysis relative to the knowledge we already had? (vi) the PCA and k-means, why are they included in the methods section? You already describe their results here. (vii) Data preparation: why is the data preparation phase described only here, and not at the start of the methods section? (viii) I appreciate the information and reasons why the drug variables are excluded (P36). We should be able to locate somewhere the list of 196 variables that are included, together with a description (P36). (ix) Imputation (how) was not clear to me. (x) Transformation is not clear to me either; which variables are transformed and how? (xi) Why having a second imputation section? Preferably, all imputation related text is presented together.

(xii) Feature selection. Why is this still 'method description'? For me, this section is one of the most vital ones in the paper. I understand and appreciate the exclusion of the variables closely related to frailty. Rather than listing labels that have no meaning, I prefer discussing (and listing in Table 2) the variables with respect to their meaning.

How do you define accuracy? (P47)

(Very) briefly explain 10-fold cross validation (P49)

"This classifier shows an extra-ordinary performance in the task of detecting non-frail patients" (P50). Please, first discuss (and add to the table) the probability a case is identified as frail using each method; this may explain differences across methods.

How do we interpret the results of for example the SVM, that is, what causes its good performance? If you do not know, just add this (I can imagine you do not know).

First line can be omitted (P52)

What result leads to the first conclusion? (P52,3-7)

Second paragraph - is it needed?
Third paragraph - yes, but was the cluster analysis and PCA needed? That is, were its results used for the learning algorithms? What was the connection between the two?

Third and seventh paragraph overlap - avoid that.

Which phase? (P53, 4)

In other studies, gender indeed predicted frailty, but the effect was not that large. You may want to cite those studies.

Limitations? - add them to discussion/conclusions.

Make sure that your figure captions describe all contents of the figures. For instance, Figure 2 (what are dimensions and percentages?), Figure 4 (unclear, statistics of what?), etc (check all).

Table 1: Feature name may be deleted, i.e., verbal description of feature is more informative. Same holds for Table 2.

Table 4: Define accuracy. Is it P("frail"| frail) + P("non-frail"| non-frail)? If that is the definition, the rule "everybody is non-frail" has a very high accuracy. Please provide info on how many in the sample were identified as frail.

Figure 1 is not optimally clear: Are the UPM patients in the middle the same as to the left? How many "more patients" were added in the middle? Similarly, what about the Aber patients?

Figure 3: what do the axes represent? Similar in other Figures!

Why are the figures of p38gpt included, and what they represent; please change p38gpt to a meaningful label.

Is the first figure needed with MICE?

Fonts of many figures are too small for me to read.

**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.

Unable to assess

**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 am able to assess the statistics

**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 for this paper.
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