Author’s response to reviews

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

Authors:

Andreas Hassler (a.hassler@hci-kdd.org)

Ernestina Menasalvas (emenasalvas@fi.upm.es)

Francisco Garcia-Garcia (franjogarcia@telefonica.net)

Leocadio Rodriguez-Manas (leocadio.rodriguez@madrid.org)

Andreas Holzinger (a.holzinger@tugraz.at; andreas.holzinger@medunigraz.at)

Version: 1 Date: 13 Dec 2017

Author’s response to reviews:

Dear Reviewers,

We really appreciate your in-depth comments, corrections, and consideration of the manuscript. We have revised the manuscript as suggested. A point by point response to your comments is provided below.

General comments:

Marcel A L M Van Assen (Reviewer 1): I 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.
Response: Thank you very much for your constructive comments, we benefited a lot from them. We agree that the paper is too long and at some points needed information is missing. We have now fully addressed all issues you have pointed out very carefully in the manuscript, and list our response point by point in this response letter.

Major comments:

Abstract: Delete 2nd sentence (always important topic)

→ As suggested by the reviewer this sentence has been deleted.

Fourth sentence: rephrase or delete., because postponing bad outcomes may even increase health costs if people live even longer.

→ This sentence has been deleted.

Background section is (too) long.

→ We agree, therefore the description of the study by which the data was obtained has been moved to the Abstract-Section Methods (Data).

Methods section is unclear: what parameters (list kind of variables)?

→ For more clarity it has been described with more detail.

CRISP-DM?

→ It is short for cross-industry standard process for data mining. For better understanding the name in full form has been stated.

data understanding and pre-processing phase?

→ These terms refer to the CRISP-DM. The part has been rephrased for better understanding.

EHR?

→ It is short for electronic health record. The name in its full form has been stated.

Results: first part is repetition, second part - what is the actual contribution (as I do not see anything particularly new)?

→ The reviewer is right. Thus, this part of the Abstract has been rephrased and the actual contribution has been stated more clearly.
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.

→ English has been checked by a native speaker and the background sections has been divided – as suggested by the reviewer - into multiple paragraphs.

P3, 23: Make clear the design, i.e., list the variables that were assessed and when. This is vital information.

→ The description of the available data and the description of the study from which it has been obtained, has been moved to the methods section where it is explained more exhaustively.

Objectives: "This model could …" does not belong to the objective.

→ Thank you for the comment. The sentence has been moved from the objective to the next paragraph.

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.

→ The reviewer is right. The related work section is to exhaustive and contains sometimes not directly related information.

→ Therefore, the section has been compressed and irrelevant papers have been removed.

→ The frailty section has been extended by papers from the medical community. Known risk factors and also protective factors have been listed.

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).

→ The C-statistic (sometimes called the “concordance” statistic or C-index) is a measure of goodness of fit for binary outcomes in a logistic regression model. The concordance statistic is equal to the area under a ROC curve. Because of that, instead of “c-statistics” “values for
the Area Under the Curve (AUC)” has been put there for better understanding. And all the performance measures are explained later in the modelling section.

→ In general, mentioned variables are described more exhaustively.

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)?

→ @ (i): Thank you for the comment. A more exhaustive description of the available variables has been given.

→ @ (ii): For better understanding the bullets of the list have been described with more detail and additionally the type of data has been stated (numeric, binary, etc.).

→ @ (iii): More content to the part where the findings of Bayles et al. are described has been added. The frailty subsection has been extended with more literature.

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.

→ @ (i): CERs and mProbes are used for biomarker discovery in bioinformatics. They were applied to gene expression datasets which makes them irrelevant for this paper. Therefore, this part has been removed.

→ @ (ii): The reviewer is completely right. As this section only contains not directly relevant papers, it has been deleted completely.

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"? ✓ (iii) what are "anthropometric data"? (iv) definition of variables: is there a list, possibly in supplementary materials, which explains in detail how each of your features is assessed?
→ (i): “CRISP-DM establishes the main tasks but does not establish a life cycle.” — life cycle was referring to the CRISP-DM development, to limit confusion “life cycle” has been replaced with “specific procedure” for better understanding.

→ The Business Understanding section has been removed and the relevant content has been moved – as suggested by the reviewer – to the introduction section.

→ (ii): The method section was structured according to the CRISP-DM phases. We understand that this is confusing and we therefore renamed it – according to reviewer suggestion- to “Data”.

→ (iii): The definition of "anthropometric data" has been added (mass and length of body segments).

→ (iv): The used features/variables are now in detail explained in the provided Additional File 1.

(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): Here (P28-30) too much irrelevant information about the data cleaning process is given.

→ All this part was referring to the cleaning process of the data and consequently it is irrelevant if you focus on the final analysis. Hence, it has been removed.

→ (vi): The reviewer is right. The PCA has only been used for data exploration and has not been used for other purposes as to analyze the relationships between the variables and the frailty syndrome. Therefore, for simplicity the approach of Wartner et al. [2016] (2-dimensional PCA plot) has been used.

(vii) 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? (viii) the PCA and k-means, why are they included in the methods section? You already describe their results here.

→ (vii): We have put this part only because it followed in the original process prior to performing the predictive modelling. As it has been shown and pointed out by the reviewer, with this data set the clustering process did not give additional knowledge to that obtained by the PCA analysis. Consequently, the clustering section has been removed. To reflect this, a sentence has been added to the discussion.

→ (viii): The reviewer is right. The section contains results and methods. Now, the Methods section only contains the methodology and the data set description. Additionally, a Results section has been added that contains the major findings of each phase of the CRISP-DM model.

(ix) Data preparation: why is the data preparation phase described only here, and not at the start of the methods section? (x) 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). (xi) Imputation (how) was not clear to me. (xii) Transformation is not clear to me either; which variables are transformed and how? (xiii) Why having a second imputation section? Preferably, all imputation related text is presented together.

→ (ix): As it has been previously said, the data preparation is the 2nd stage of the results section.

Data preparation is described at this state because it has been mentioned at the very beginning the analysis follows the CRISP-DM methodology and consequently preparation goes after data understanding. Because we were following the structure of the CRISP-DM phases (Business understanding □ Data understanding □ Data preparation □ …)

→ (x): Thank you for the comment. The used variables can now be seen in the Additional File 1, which provides information about all the available variables.

→ (xii): After analyzing it again we realized this part belongs to the cleaning activities, which as mentioned earlier, have been removed.

→ (xi) & (xiii): The 2nd imputation was required by the doctors but after analyzing it, we found that it is confusing so we have removed it.

(xiv) 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.
As has been mentioned the section has been removed and Feature Selection is now a subsection of Results.

As suggested by the reviewer, the labels have been removed and replaced by descriptions of their meaning.

How do you define accuracy? (P47)

A description of the used performance measures has been added.

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

Thank you for the comment. An explanation has been added.

"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.

We were referring to the other results, therefor this sentence has been removed.

Table 4 contains a comparison of the performance of the different 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).

The performance measures have been explained in subsection Performance Measures. So, we hope that this now clarifies the performances measures that have been used in the paper.

First line can be omitted (P52)

Following the reviewer suggestion, the line has been omitted.

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.

→ Thank you for the comments. As this part was not clear at all, the complete discussion section has been redone.

Limitations? - add them to discussion/conclusions.

→ As suggested, the limitations have been added to the discussion.

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).

→ Additional information for each figure has been added.

→ As the reviewer is aware the BMC Bioinformatics Journal has a limitation on the caption, nevertheless we have included as much information as possible.

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

→ Feature names have been removed and replaced with verbal descriptions.

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.

→ A definition/explanation of the used performance measures (subsection Performance Measures) has been added.

→ The number of frail/pre-frail and non-frail subjects has been added.

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? ×

→ This figure has been removed as it is irrelevant for this research. The used data set that has been described in the section Methods.

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

→ Thank you for the comment. All the figures have been reviewed and a description has been added to explain their content. In particular, for Figure 3 a description of the axes has been added (in the new manuscript it is Figure 2).
Why are the figures of p38gpt included, and what they represent; please change p38gpt to a meaningful label.

→ The figure was included just to show how the process of visualization was performed and how the variables can be assessed regarding outlierness, but it is true that it doesn’t really add relevant content and consequently, we have removed it from the paper.

Is the first figure needed with MICE?

→ Once again, the reviewer is right that the figure does not add relevant content, therefore it has been removed.

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

→ Thank you for the comment. Fonts of the figures have been made bigger.

Dustin French (Reviewer 2): I agree with your idea of using EHR info for prediction analytics.

1. Streamline Intro and Lit Review to reflect your contribution in geriatrics. I would not get into drug surveillance and other disease predictors outside of your area of emphasis.

→ Thank you for the comment. According to the suggestion we have condensed the introduction and the related work section. The papers of other domains/diseases have been excluded and only relevant information gained by those has been added.

2. There are many predictors in this model. From a stats point of view this presents problems of Multicollinearity and degrees of freedom problems

→ We agree with the reviewer but multicollinearity is described as a problem only if you are interested in the effects of individual predictors. It does not affect the predictive power of the models. Therefore, the overall performance of the build models is not affected by the Multicollinearity. However, we added information to the paper regarding this issue.

→ In the Machine learning community feature selection is used as a tool to remove variables and therefore only the more predictive ones remain.
3. Many of these variable are part of AGS guidelines for fall prevention. They are also RAP triggers in the MDS for LTC. You may want to use these clinical tools/guidelines as justification for variables.

→ This was a clinical trial which has already been done. The variables have been decided by the designers of the Toledo study. We follow the criteria the clinicians established. As no other variables were available and there was no possibility to obtain other ones.

4. The number of figures and info is overwhelming. Again can you condense what you have? Think about what you are adding to the literature. There have been other papers prediction 30day readmissions, falls in the elderly, LTC,etc. I did not see these here but you should think about adding them in.

→ Thank you for the comment. The complete paper has been restructured, sections have been condensed, irrelevant information and figures have been removed in an attempt to make the paper more readable. Regarding the literature section, it has been also reviewed, papers that were irrelevant have been removed and references suggested by the reviewers have been included. However, papers about LTC have not been included as we think they are out of the focus of the paper.

Thank you again for all your suggestions and comments.