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: 2 Date: 04 Jul 2018

Author’s response to reviews:

Note: I'll upload this responses also to the system for easier reading.

Dear Reviewer,

We really appreciate your in-depth comments, corrections, and consideration of the manuscript. We have revised the manuscript as suggested. The paper has been condensed and restructured and we hope it is easier to read and the information is presented in a more suitable way. A point by point response to your comments is provided below.

Reviewer reports:

Marcel A L M Van Assen (Reviewer 1): Please include all comments for the authors in this box rather than uploading your report as an attachment. Please only upload as attachments annotated versions of manuscripts, graphs, supporting materials or other aspects of your report which cannot be included in a text format.

Please overwrite this text when adding your comments to the authors.

We assume this text is for the other reviewers.
Bonnie Westra (Reviewer 3): The authors address a topic of importance and apply data mining techniques with an emphasis on data preparation. I am including my comments as an uploaded document as it is difficult to maintain formatting to understand the comments. While the authors addressed previous reviewer comments, there remains multiple issues of concern.

The paper continues to be longer than one would expect for a journal. There continues to be considerable redundancy in sections that could be combined and sections that could be reduced or removed as noted in my comments submitted.

We have analyzed redundancy and the structured as suggested. As a main consequence:

- Parts of the results which in fact were methods/discussion/limitations have been moved to the respective section
- CRISP-DM has been removed
- PCA has been removed
- Background, Related Work and Discussion section has been rewritten and re-structured
- Limitations subsection has been introduced

The focus of the paper is on predicting the frailty syndrome, yet a definition is lacking until p. 12 and then the future work identifies frailty as a disease - clear, consistent use of this term is needed and should be introduced very early in the paper.

The reviewer is completely right. Even that we had a definition in the introduction, the state of the art section did not include the definition. Consequently, the definition of the frailty syndrome has been included in the Related Work section (page 5) and further, frailty has been introduced in the Background (page 3) and a reference to the exact definition in the Related Work section has been provided.

The background section has been reduced, but it continues to lack integration based on the main arguments for why a study is included - the "so what" is missing to justify the need for the current study based on previous research. This section should be about key points supported by the literature to show how you are building on previous studies and the gap that this investigation will address or to support what the investigators will be doing.
Thank you for your input.

References have been again checked according to the reviewer suggestions and consequently:

• For each reference a “so what” an analysis has been added

• The references in question, which are not directly related to frailty (risk for falls, long-term survival etc.) have been removed as they were not really relevant for the focus of the paper.

The most significant problem is still not addressed based on a previous reviewer suggestion a traditional format of intro, methods, results, conclusion/discussion) - these headings were used - however, results included methods, findings (sometimes limited) and interpretation. Often the results starts with additional comments.

The structure of the paper has been revised and the following changes have been made according to the suggestions:

• Background content is now in the section of the background

• Method section now includes the definition and description of all the methods used

• Results: the part that was in fact methods has been removed from there

• Discussion includes the main discussion on the results obtained

• Limitations subsection has been added

PCA - it is not clear how this is useful for the rest of the paper. After describing this section - it doesn't seem to inform further analysis.

The reviewer is right that the PCA section does not contribute to the presented research, as it was just used for explorative analysis, and consequently it has been removed.

Please see uploaded comments for more details with additional comments

Thank you for your detailed and also very helpful suggestions. They have been addressed and please find the response also in this document.
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 reviewer is right that the presented research does in fact not work with typical electronic health records but structured medical study data consequently:

- EHR has been removed from the title
- In the rest of the document EHR mentions have been properly eliminated

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 main reason of the fusion of the pre-frail and frail status was to try to analyze if non-frail patients differ from those in early stages of the frailty syndrome. Moreover, this work can be considered as the first attempt to predict frailty not using the known variables by Fried et al. 2001. That is the main reason for the merge into 2 classes. Once results are validated, following stages of the research should go (as we have stated in future lines) in the direction of predicting the three stages of the Fried score (multiclass problem).

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.

In fact, now in page 3 and 5 we have included the definition of Frailty and we have stated that one of the goals is finding predictors for frailty and consequently we tried finding early predictors that distinguish non-frail individuals from those being even in the early stages of frailty.

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

In the created „Limitations“-subsection of „Discussion“ we have addressed this issue and further pointed at it in the “Future Work” section.

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.

We are really thankful for the comment. In fact, this was the main goal of the paper and consequently this has been clearly stated in the goal section, as suggested.
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 Niaksu, but it is not explained to which extent 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?

In fact, the paper follows the CRISP-DM process model but it is true that the whole definition of CRISP-DM is not required and consequently according to the suggestion this has been removed

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

In the method section several imputation methods have been used. However, none of them helps with the quality of the model as the main problem is the quality of the data. In any case in the section limitations both the improvement in imputation methods together with the quality of data have been included.

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.

According to the reviewer the paper now includes a section of limitations and goals have been clarified.
Provided Additional Document: MIDM-D-17-00311_R1 – new review

Purpose of the study

The purpose of the study is identifying risk factors for discovery of the frailty syndrome, but the key words do not include frailty syndrome – recommend adding

As has been suggested by the reviewer, the key word “frailty syndrome” has been added.

The goal section has been improved according to the suggestions and in fact now it is written frailty syndrome or simply frailty.

Background

• Previous review indicated need for integration of literature to support major points. While this likely improved, there is more work that is needed.
  o P. 7 - Ramakrishnan et al. said something

According to the suggestions this reference has been deleted as it is not really relevant to the proposed research

o P. 8 - Goldstein et al. – lots of critique – but then so what? How does that support the study other than there is room for improvement

As in the previous case, this reference has been eliminated in this version.

o P. 9 – Frailty – these studies are discussed as separate paragraphs and unclear why they are there – they don’t include a definition of frailty nor is it clear what the point is
  ▪ Tinetti et al.
  ▪ Friedman et al.
  ▪ Swindell et al.

o Baylis et al. – relates immune-endocrine axis and frailty - Fried score for frailty assessment introduced but no clear idea of what this is

As part of the improvement of the background section, the definitions regarding frailty have been “glued” together and those not relevant, have been removed and more relevant references have been provided.
Methods

CRISP DM is a framework for the steps in the data mining process - there is a lack of methods used in this study outlined in the methods, rather they are noted under results.

As CRISP-DM is not central for this research, for clarity it has been removed.

Data Source – In the methods section – the data source was secondary use of data from a previous study, in other sections, the discussion notes that EHR data can be used – so it is somewhat confusing.

Abstract - a methodology has been described, which makes it possible to explore and to make use of the incomplete, inhomogeneous EHR data and further to identify potential predictors. (In the methodology – it is stated that this study was secondary use of data from the Toledo study not EHR data)

P. 5 – under related work – demonstrate how a data mining approach with EHR data can be helpful

Conclusion “In this paper the feasibility of applying data mining techniques in order to extract models for frailty prediction using EHRs from patients some of which are frail, has been analyzed and a methodology has been presented.”

As it has been already explained in the general comments the reviewer is right and consequently we have removed the reference to EHR, firstly in the title and later in all those parts in which it was mentioned. It has been stated clearly what kind of data has been used for the research.

Definition of Variables

Typically, this section is introduced with definition of predictor variables and then outcome.

According to the suggestion, at the very beginning of the definition of the variables we have put a sentence to state that the variable to be predicted is frail and that the rest of the variables will be used as predictors.

This sentence is unnecessary – “As it has just been explained, patients are described by a set of 284 variables.”

The sentence has been removed.
p. 14 – ADL – noting there are different tests – the issue here is what was used – ADL is defined as xxx and operationalized using xxx test. In IADL – it was noted that the sum score was 0 – 8, but this wasn’t clear for ADL

This part of the paper has been clarified, and now for each different test the range of the values/scores is stated, and further also which kind ADL test was used (The Katz Index of Independence in Activities of Daily Living).

Definition of outcome

• Since this paper is about predicting the frailty syndrome, this should be defined early in the paper so the reader knows what is predicted in the study

• P. 3 – prevalence and risk associated with frailty identified- no definition is provided. It is synonymous with disability, comorbidity and other characteristics – but no definition nor how it might be similar to different from these concepts. If frailty is the outcome – it has to be clear to predict it.

• P. 4 - The prevalence of frailty in people 65+ ranges from 7% to 16.3%, increasing with age, and it is the main risk factor for disability. – still no definition of what this is and no references for these statistics

• P. 12 - The diagnosis of the frailty syndrome was based on the Fried criteria (weakness, low speed, low physical activity, exhaustion, and weight loss) – It is surprising that it took until p. 12 to know what this critical concepts means.

• P. 39 – Future work – “This work has contributed towards the possibility of obtaining predictive models that can anticipate the onset of a disease. In particular, the problem of frailty has been analyzed in this paper.” – This doesn’t logically follow as frailty was not defined as a disease.

Thank you for your input, this issue has been clarified and additionally to an early explanation on what the frailty syndrome is and how it is defined, coherent wording in the document has been used where frailty is not called a disease but a syndrome.

Results

• Results typically begin with a description of the population, then results are provided for each aim/ objective.
In the new structure of the paper the results only contain the results to solve this problem.

- Visualization and analysis of values

The Result section now only contains the visualization of the results.

- Most of the results section reports methods and not results - there are no results reported in this section

According to the suggestion of the reviewer, for clarity the structure of the paper has been analyzed and content has been reassigned to the corresponding sections. Now all the methods are under the method section and the result section only contains the results of the analysis.

- Additional file 1 is the data dictionary - no reference to figures or tables with results

- The Additional File 1 contains the complete variables description. The results are all shown in the paper. To avoid confusion, we have removed references to the Additional File 1 in the results section and we have explained in section “Data exploration and quality assessment” what this file contains.

- Outlier and missing value assessment

- this section lacks specific results, includes opinions and doesn't provide any information about how missing data was handled in this study (which should be in the methods section)

→ According to the suggestion, the subsection has been rewritten and shortened.

- Ontology-guided PCA

- The PCA should have been in the methods section, then only results reported in this section. How is PCA a data mining approach? This is traditional statistics

- Why would this be done in addition to data mining and visualization?

- MD recommended variables: education status, income, body mass index (BMI, self-derived), Geriatric Depression Scale (GDS) score, total number of comorbidities (self-derived), Mobility Score (self derived), gender and polypharmacy.

- Findings – what do the numbers mean along the axes?

- It appears that the PCA did not inform which variables to keep in the data preparation section – therefore it is unclear why this was done

→ Once again, the reviewer is right and the PCA part does not contribute to the main goal of this research, so for more clarity we have removed it.
• Why does data preparation follow the previous sections in results. Typically, data
preparation is the first step after data selection –

- The paper has been restructured. The used order of methods/approaches originates from
the used CRISP-DM approach for the data mining process (Business Understanding <-> Data
Understanding -> Data Preparation <-> Modeling -> Evaluation -> Deployment). The previous
section corresponds to the task “Data Understanding”. In the new version of the manuscript Data
preparation only is presented once before data modeling phase.

• Data file preparation – first paragraph is not essential in a publication

- Thank you for your input, once again following the suggestion the paragraph has been
removed.

- Feature creation – one example was given – not clear if other features besides BMI were
created

- This has been clarified and the other derived features have been mentioned and
explained.

- Earlier – missing data was identified – in this section the methods were discussed – it
would be clearer to combine and only point out once. How missing data were handled is a
methods issue and the results should be included in the results section – opinions in the
discussion section

- As the paper has been restructured, rewritten and condensed, we hope the new
manuscript is less redundant, better structured and more clear.

• Dimensionality reduction – feature selection

- There are multiple sections addressing this issue – PCA, MD opinions, then a general
sentence – but feature selection is under another heading. It seems combining these together in
logical sequence in one section would be clearer to the reader

- Methods, results, interpretation – these are all combined in the section called “Results”

→ The overall structure has been redone and all the contents have been moved to the right
section according to the suggestions.

• Modeling and evaluation

- Methods, results, interpretation – these are all combined in the section called “Results”
The overall structure has been redone and all the contents have been moved to the right section.

- Multiple methods used – it would be helpful to justify why these were selected
  → A justification has been provided aiming to clarify this.

- Major section on data preparation
  → This section has been moved to the general data preparation part.

- P. 31 – first time the frequency for the outcome introduced – it would be helpful to introduce this earlier in the results. Also it would be helpful to indicate that the same number of frail vs nonfrail were included in each sample
  → Thank you for the input. The frequency for the outcome has been placed earlier in the paper (Definition of Variables). Moreover, a sentence has been added which clarifies the goal of the stratification procedure (maintaining initial class balance).

- Model performance – good explanation of the measures
  → Thank you for the comment, however at this stage we have contradictory suggestions, in fact another reviewer has suggested to remove this part, as those are commonly used performance measures. As the overall concern with the paper is the length, it has been decided to remove the detailed explanation of the measures and just shortly mention which ones have been used and a reference has been added for the interested reader to find further details for the measures.

- Analysis of business goals – starts with an opinion “it can be said that they have been found”
  → Thank you for your input once again according to the suggestion the sentence has been removed.

General writing problems

- References are represented on [?] and no reference list included
  o This was an issue with the submitted latex file which was written in Overleaf and apparently not compileable in this form at BMC Bioinformatics.
  o We tried to handle this issue and the references should be displayed correctly now.

- p. 4 – new paragraph, with no indenting
o Issue has been resolved.

• Multiple times single sentence paragraphs

o p. 4 lines 37 – 53

o p. 5 lines 40 – 51

o p. 13 – lines 33-37

o p. 37 lines 18-22

o multiple instances in future work

We have reworded and joined the sentences into paragraphs.

• Spell out acronyms first time of use i.e. p.e IET-HEALTH, EU

Issue has been resolved.

• Unclear meaning or unsupported statements or reason for including information

o p.2 - identified collection of tools (computational methods) as an essential instrument – unsure what this means

We have reworded it for clarity.

o [data mining] It seems to be the current practice that there is a ”black box”. – this is not correct – when reading this paragraph it seems that it is about there are different methods – some of which are black boxes and some are glass boxes – rather than the current method is black boxes

According to the reviewer suggestions this paragraph has been rewritten and open questions and issues have been clarified/solved.

o DM reference model in the medical domain and introduces 38 new generic tasks as extension of the model. – what is the relevance of this statement as it doesn’t seem to influence later methods or isn’t clearly linked.

Has been deleted altogether with the CRISP-DM subsections.

o P. 10 - The through forward search identified 13 features are listed here: - what does this mean?
Forward search is a feature selection approach which has been used by the authors in order to detect suitable predictors. However, this paragraph has been removed in the rewriting and shortening process.

- Headings for sections – then the content doesn’t match what is anticipated under that heading

  o P. 3 - Understanding of the Frailty Problem and Translation to Data Analytics – under this heading the first paragraph is about a project

The section has been restructured for more clarity and headings have been changed.

This particular heading has been removed.

  o P. 6 - Many issues regarding privacy, ownership, regulatory requirements, ethical, political, technical and social aspects of health data still remain. This sentence doesn’t tell the reader what the paragraph is about as the paragraph goes on to discuss many other issues i.e. heterogeneity of volume of data, need for standardized procedures, and then paragraph ends in data warehousing

The sentence has been connected to the previous paragraph and the content has been adapted appropriately. The part about data warehousing is definitely out of the scope of this work and hence, has been deleted.

- Redundancy is discussing an issue in multiple places

  o P. 7 - CRISP-DM in the Medical Domain

→ CRISP-DM has been deleted. Paper has been checked for redundancy.

→ We hope the new structure has minimized/eliminated the redundancy.

- Need to look at headings used for the journal

  o i) Data Visualization and Analysis of Values

→ Thank you for your input. These headings have been removed or adapted according to the guidelines.

- P. 20, line 23 “In the next chapter the aforementioned issues will be treated.” Unclear why a section in an article is titled chapter

  o Thank you, this issue has been resolved and the reference to a chapter is eliminated.
I am happy to read that the authors tried to follow upon the suggestions of the reviewers as best as they could. The paper improved a lot. Although the paper is no book anymore, it is also not yet a paper; it can be improved a lot, still. What deserves most attention is the results section, which is way too long (> 20 pages) with dozens of subheadings and very hard to read. In the major comments section I provide some ways to shorten it a lot, but it was difficult for me to do so as it is so far away from what I see as the ideal results section. I still believe in the value of this research though.

Thank you! Everything has been completely restructured, parts have been rewritten and condensed, thanks to the suggestion of you and the other reviewer. The paper is now considerably shorter and we hope it is more readable now.

For the first part, before the results section, I also provide many minor comments. In a future revision I hope to provide also minor comments on results and conclusion, as it is now difficult for me to do so because I get lost in the results section. I therefore would expect two other rounds of revision; one to streamline and condense the paper even more, and the second for the details. I hope the authors are able to get through the next revision satisfactorily.

According to the suggestions of the reviewers we have restructured the paper and now most of the content that was under results is now under methods and we have simply left the results obtained running the algorithms for the results section.

Major comments

Most of my comments to the first version of this paper concerned writing (grammar and English), length, structure, and information (a lot less needed, and some missing). In all respects the paper improved. However, the paper is still much too long. Its structure also should be improved, as I still find information at unusual places. Grammar and English seem to be fine (although I am not a native). Specific suggestions how to improve each section of the paper can be found at minor comments below.

Thank you very much for your input.

I do not provide very specific comments to the results section, as this is still far from what I see as a good results section. It seems to be (about) 22 pages long, contains tens of headings, and includes a lot of information that imo are no results but methods. More than half of these 22 pages can be removed/condensed.
Thanks for the comments. According to the suggestions the paper is now restructured and most of the information that was under results is now under the method section.

Most of the content as you and the other reviewer noted, does not belong to the result section. Therefore, after re-structuring everything it was rewritten and is now much shorter.

The section “i) data visualization’ does not present results, and is more a methods section. The same holds for “ii) outlier….”. I do not see results. In a methods section? “iii) PCA” is ok, but i)-iii) can (should?) be condensed considerably. For the PCA, please mention the number of variables entered and the number of cases. Consider using Parallel Analysis to check if you selected the adequate number of components. Why include frailty in PCA as well? I do not see reasons for that, I thought PCA was about reducing number of predictors? Please, use less text explaining PCA, and remove subheadings.

The reviewer is right and consequently we have removed all the parts concerning the PCA analysis, as it just has been used for exploratory analysis and does not contribute to the main work of this research.

Data preparation is often not considered as “results”. Why data preparation AFTER data exploration? If you exclude some features, then you do not explore these, do you? Anyway, data preparation section can be condensed into a much shorter text. Only from p28 (line 25) I get the feeling that we read something vital. All details about the analyses can be in supplements, main storyline in the paper.

We did follow the CRISP-DM approach for the data mining process (Business Understanding -> Data Understanding -> Data Preparation -> …). Therefore, this order of steps/methods has been chosen. Data preparation has been moved to methods. We hope in the new manuscript everything is more transparent.

Modeling and evaluation section is imo important, and I do not see reasons to condense the classification model settings section. I would remove data set preparation and just tell the reader that you applied the models to standardized variables or min-max normalized variables, where you give the formula for the latter normalization. This is just a few sentences.

According to the suggestion the text has been shortened and were not essential removed.

Figure 7 can be removed (does not help). Model evaluation section can be shortened and incorporated in the model performance section.

Following the suggestion, figure 7 has been removed.

Performance measures can also be reduced a lot, as most measures are well-known. One short paragraph is enough, and no formulas are needed.
Even that another reviewer wanted to keep this section, we considered this but then decided for the sake of a shorter paper (which is the main goal), to remove the definitions and to provide references for the interested reader.

Model performance is important. “Business goal compliance”, where does that come from? Why not just delete these words and continue interpreting the best predictors?

According to the suggestion, these words have been deleted.

Consider rewriting the discussion section as well. Relate it to the intro. Goal is to predict frailty using data mining, right? Start with the goal in the first paragraph, and possibly mention the data set you use. This all in one (the first paragraph). Then mention and discuss your main findings. I suggest you discuss your findings along two lines, and sequentially; either first your findings concerning methods and then concerning predicting frailty (seems most logical to me), or the other way around.

Goals have been clarified at the very beginning and the discussion has been rewritten trying to differentiate goals achieved. Also, a part for limitations has been added. We hope that the discussion section now is clearer.

Minor comments

Why a question mark and not a reference? (2:27) Your paper contains many question marks where the reader expects references.

- This was an issue with the submitted latex file which was written in Overleaf and apparently not compileable in this form at BMC Bioinformatics.

- We tried to handle this issue and the references should be displayed correctly now.

Definition of frailty is long, including words that are not part of the definition (3:14-22).

Thank you for your input, this issue has been resolved and the frailty definition has been checked and shortened

“Furthermore, this would…”; would □ may (3: 33)

Thank you. This has been changed according to your suggestion.

“Understand of the frailty…” section. The second paragraph of “Understanding of the frailty…” can be removed/deleted, as most of this info should be in intro (the avoiding disability part). The sentence about the goal of FACET can be incorporated in the first paragraph. Third paragraph about frailty can also in previous paragraph, because it relates to definition of frailty. The fourth and fifth are about FACET, and are fine (and possible can be incorporated in the first). Objective
of the paper should be in intro (sixth paragraph, 4: 19-22). After these changes, you’ll see that a more natural label for this section should be something with FACET. But, why keep this large section on FACET? Is it necessary? I would skip it entirely, and mention in the intro in one or two sentences that this project is carried out in the context of FACET. The Objectives paragraph is then a natural end of the introduction.

Thanks for the comment. According to the suggestion this part has been rewritten and re-structured.

The parts related to FACET has been removed, as suggested.

Helpful for what? (5: 44)

It has been clarified “The main focus of this paper lies on demonstrating how a data mining approach for health data can be helpful for discovering new potential predictors and for building predictive models”

The first two paragraphs of “related work” can be combined into one, using only one sentence what you will review (essentially skipping the first paragraph). Try to shorten your text everywhere, deleting unnecessary text.

Thanks again for the suggestion. We have rewritten the paragraph and shortened text in the whole paper.

“Data mining in the medical domain” section. This section has four paragraphs. First is ok. The second is ok but can be organized better (“these methods” which?; The Prokosch part is long, and partly repeats the first part of the paragraph). Third paragraph can be skipped or incorporated in the first one? Fourth is fine in spirit, but long; please only include info that is necessary.

We have rewritten the paragraphs and removed unnecessary content.

“CRISP-DM” section. It has three paragraphs. In the first, you can delete third and fourth sentence. Relevance of Belazzi to CRISP-DM is unclear. The relation of the second section to CRISP-DM is completely unclear. And what is the purpose of the third paragraph? After reading this section on CRISP-DM I am at a loss, and have no clue what CRISP-DM actually is. This section needs to be revised completely such that we know what CRISP-DM is and what we should do with it.

We have removed the part of CRISP-DM, in fact as the reviewer suggests, it does not contribute to the content of the paper.

“Frailty” section. This section has four paragraphs. The first two paragraphs are not on frailty but on falls and fear of falling. I expect previous literature on frailty and its relevance to your project.
The third and fourth paragraph are both more than a page long = much too long! They are on survival; what is their relation with frailty? This section needs to be revised too.

Thank you for your input. According to your suggestions we have rewritten the whole section and removed content which is not directly related (risk/fear of falls, survival, etc.).

“CRIPS-DM” in methods; one sentence “we use…” is enough, heading can be deleted.

We changed this part as suggested. (CRIPS-DM has been removed)

“Data”. Good! Incorporate first in second paragraph, where the first paragraph can be summarized into one sentence which is the start of the whole section (consisting of one paragraph) (“We use the data of .”).

It has been restructured & rewritten according to the suggestion.

“Definition of the data sets” can be added to the previous section.

Has been done according to the suggestion.

“Definition of the variables”. Paragraphs of one sentence are not needed. Description of scales is good.

It has been restructured according to the suggestion.

Section i) Data visualization: why does it include one-sentence paragraphs?

This issue has been resolved.

Delete “all following … components” (18: 5-10)

The paragraph you are referring to has been removed in the restructured and rewriting process.

Thank you again for all your suggestions and comments,
we very much appreciate your reviews

Sincerely,

Hassler, Menasalvas, García-García, Rodríguez-Mañas and Holzinger