Author’s response to reviews

Title: Languages for different health information readers: Multitrait-multimethod content analysis of Cochrane systematic reviews textual summary formats

Authors:

Jasna Karačić (jasnakaracic@hotmail.com)
Pierpaolo Dondio (pierpaolo.dondio@dit.ie)
Ivan Buljan (ivan.buljan@mefst.hr)
Darko Hren (dhren@ffst.hr)
Ana Marusic (ana.marusic@mefst.hr)

Version: 1 Date: 22 Jan 2019

Author’s response to reviews:

Editor’s Comments:

- Please provide a "Conclusion" section.
ANSWER: Corrected.

- Please move Abbreviations after the Conclusions section.
ANSWER: Corrected.

- Please provide Acknowledgements section in the Declarations
ANSWER: Corrected.

- Please provide contributions of Darko Hren in Authors Contributions section (use initials only).
ANSWER: Corrected.

- Please remove Figure titles in the figure files
William Speier (Reviewer 1):

The authors present an analysis of the readability, tone, and style of different formats of scientific text from the Cochrane web site in several different languages. The topic of readability is vital for the understanding of medical literature and has been shown to affect medical decision making. In particular, I appreciate the inclusion of translations of the clinical text as most readability assessments that I have seen focus on a single language (usually English). The analyses performed by the authors are conducted well; the results are interesting and generally agree with previous studies that clinical and scientific text, including "plain language" summaries, is generally written at too high a level for the general population to understand. While I believe the importance of the manuscript is high and the analyses are appropriate, there are a few areas that I believe are currently lacking which should be addressed.

1. The background section should generally be expanded. While the authors include some good references of previous studies that have tested readability of health literature, the background does not currently mention these studies explicitly or how the current study builds off of previous findings. The authors also do not fully explain what Cochrane is and the reasoning behind using it. For context, it would be helpful to have a brief review of some other efforts to make health information available to the public, describe how Cochrane compares to the other efforts, and say why it was chosen for the current analysis. Mentioning who performed the translations would also be helpful information as that can affect the readability and accuracy of the translations. Finally, the formula for the SMOG index is not fully explained and the authors do not discuss the other methods for readability (e.g., Flesch-Kincaid) and why SMOG was chosen for this analysis.

ANSWER: We thank the reviewer for these comments. We wanted to keep our manuscript succinct and brief for the professional reader, but we agree that the Introduction section would benefit from details on what efforts are there available for making health information available to the public and on the Cochrane and its efforts in knowledge translation to patients. The Introduction section in the revised manuscript is now expanded and we also tried to make it
relevant and to the point. The rationale for the use of SMOG index has already been provided in the Methods section, but we also now address it briefly in the Introduction section of the revised manuscript.

2. The results are interesting, but the presentation is inconsistent and can be hard to follow. For instance, the sentiment analysis section is the only section that includes numerical results while the other sections use qualitative statements such as "did not contain strong emotional tones". The figures also do not convey the variance for the values, which can make it difficult to appreciate the differences between the different document types. Box plot might be a better way to demonstrate these distributions.

ANSWER: During writing of the manuscript we paid special attention to the presentation of data, as they are quite complex. We tried different presentation formats, including box-plot but they were quite complex because of a number of variables for each of the language analysis for each of the Cochrane systematic review summary formats. The sets of box-plots were almost as “busy” as the tables that are presented in the supplementary material to the manuscript. We decided on the radar graph presentation because that allowed us to show schematically the relationship between language characteristics for specific analysis set, as well as to show the differences in these language characteristics between different textual summary formats. We could not show the error bars (95% confidence intervals) because they were generally too small to be visible outside of the data points. The full analysis is presented in the supplementary tables. Because we present complex and novel data, we decided to leave the graphs in the revised manuscript because they better illustrate the language characteristics and differences between the textual summary formats. We also indicate in the Figure legends in the revised manuscript why error bars are not visible. We are happy to follow the editor’s instruction in this matter.

3. Finally, the authors do a good job of presenting analyses, but the hypotheses behind the analyses are not clearly stated. For instance, why is it important that sadness is the dominant emotional tone? Does that affect a reader's ability to understand or interpret the text? What is the practical significance of different languages having different reading levels? The manuscript would be stronger if it had a firmer conclusion and description of how the results of this study could be applied to improve the readability problem.

ANSWER: Our study is the first such study of the textual summary formats for systematic reviews, so it is difficult to make conclusions how different language characteristics influence a reader’s ability to understand and interpret the text. Our findings open possibilities to test hypotheses on how changing different language characteristics may impact the understanding of textual health information. This is now stated in the Discussion section of the revised manuscript.
Other comments:

1. It is interesting that Croatian and German have much lower SMOG indices than French or English. Is this because of the translations? Is there something specific about the languages that affects the readability scores?

ANSWER: We thank the reviewer for this important comment. As other languages are not often studied simultaneously, it is difficult to make valid conclusions on the differences in SMOG score for different translations of the same text. Using a very broad (sensitive) search using the terms “SMOG” and “French” we identified 19 articles in the PubMed, out of which only one compared SMOG results for English, French and Spanish. For the German language, the published studies used the same SMOG formula as for the English language. As for the Croatian, we used a special formula developed specifically for Croatian, which, as a Slavic language has extensive grammar and longer words because cases are built by adding suffixes to the nouns. We now address this issue in more detail in the Discussion section of the revised manuscript.

2. It is interesting that scientific abstracts have significantly lower analytical tone (figure 3) but significantly higher values for analytical thinking (figure 5). What is the difference between these two (other than the analyser used) and why might it have caused this apparent discrepancy?

ANSWER: Both methods identify similar characteristics of the text and the differences are probably result of different dictionaries used by the two methods. In both cases the scores are high for all types of texts showing convergence of the two methods. Concerning the discrepancy mentioned by the reviewer, we could not precisely identify its source, as Watson and LIWC are proprietary products whose dictionaries are not publicly available. Educated guess would be that LIWC scores are more informative of style and Watson on emotional tone.

3. In general, statements about "high readability" can be confusing because "high readability" means "low readability scores" due to the concept of readability being inversely related to the SMOG score. I think the authors even mix things up in at least one place ("decreasing readability" on page 8). I would suggest saying "more" or "less readable" when comparing two types of documents and "high SMOG grade" (or "high education requirement") when talking about absolute readability.

ANSWER: We thank the reviewer for the suggestions – we implemented the suggested changes throughout the revised manuscript.
4. Some acronyms are not defined (e.g., SMOG). Also, Linguistic Inquiry and Word Count is defined twice, first as (LWIC) and then as (LIWC). Both versions are used multiple times (LIWC is correct).

ANSWER: We thank the reviewer for pointing out the inconsistencies – we corrected them throughout the revised manuscript. We only left the term SMOG in the Abstract as that is the usual way of presenting the term (judged by abstracts published in PubMed).

5. The authors make a statement about "truthful information about health", but I think they mean accuracy or precision. Truthfulness usually includes connotations about honesty. I do not think the authors mean to say that changing readability will lead to deception, but rather that simplifying language may necessitate some loss of precision.

ANSWER: We thank the reviewer for this suggestion – we changed the statement accordingly in the revised manuscript.

Mike Conway (Reviewer 2):

### SUMMARY

This article describes work on analysing the readability (and other linguistic characteristics) of summary formats for Cochrane systematic reviews. The text summary types consisted of *press releases (in English), *scientific abstracts (in English), and *plain language summaries in various languages, including English, French, and Croatian. Linguistic tools used for analysis included LIWC, Stanford sentiment analysis tools, and IBM Watson Tone Analyzer, in addition to SMOG, a widely used readability metric.

The researchers found that all summary types were relatively difficult to read, especially given that the recommended reading level for health communication is 6th grade, but that plain language summaries were the easiest to read document type.

This is an interesting and generally well-written paper on an important topic (i.e. appropriate health communication), but it does have several problems (alluded to below).

### COMMENTS

p2. abstract. "In all text formats, 'sadness' was the most dominantly perceived emotional tone" - Is this not related to the subject matter (illness) rather than tone per se?
ANSWER: We agree with the reviewer that this may be the explanation for the sadness emotional tone. We are not sure that this can be addressed in the Abstract, due to the length constraints, but we address this issue in the Results section of the revised manuscript.

p2. General point: There is a comprehensive literature on hedging in scientific writing (i.e. similar to your "Tentative" category). It is probably a good idea to cite some of this work.

ANSWER: We thank the reviewer for this suggestion. We now include a citation about the tentativeness of medical research findings in the revised manuscript.

p2. Results. "All text formats had low readability, ranging from a median of 15.6…” need to specify what metric you're using here

ANSWER: We rephrased this to indicate the metric (see response to Reviewer #1)

p2. Results. "Clout" probably needs a couple of words of explanation

ANSWER: High scores on Clout reflect a perspective of high expertise and confidence, and low scores reflect tentative or humble style. This is now clarified in the Results and Discussion sections of the revised manuscript.

p2. "6th Grade level" - terminology regarding grade levels differs from country to country (e.g. in some countries kids will be older, in others younger), so it's probably important to specify which country you are referring to. I'm assuming you mean the US as the widely used readability measures typically use US grade levels.

ANSWER: Yes, the grade levels used in SMOG metrics are US grade levels, where 6th graders are pupils aged 11-12 years (see: https://www.pearsonclinical.co.uk/Sitedownloads/Miscpdfs/Gradetoage.pdf). This is now clarified in the revised manuscript.

p3. "…and that patients and general population" typo/grammar

ANSWER: Thank you, this is now corrected.
p3. "Scientific abstract is the basic form of summarising..." typo/grammar

ANSWER: Thank you, this is now corrected.

p4. Why did you use the SMOG metric only? There are a bunch of different readability metrics out there, each with its own biases. One way of addressing this involves using several metrics and taking the mean score

ANSWER: While the reviewer has a point that using several metrics and taking the mean score could reduce the bias of using a single metric, our rationale for using SMOG only was two-fold:

1) It is the most commonly used readability metric in health. For example, the Office of Cancer Communications of the US DHHS reviewed the predictive validity of 12 selected readability formulas in 1982 and chose the SMOG metric to testing readability levels of its public and patient education materials.

2) The aim of our research was not to establish the “true” value of readability for texts included in this study, but to make comparison between different textual formats of the same information (Cochrane systematic review), so that the use of a single readability measure was justified for such a comparison.

This is now cleared in the revised manuscript.

p4. The way you have chosen to describe the three levels of data analysis for the IBM tool as emotional, psychological, and writing is a bit confusing, as the first two of these could be reasonably described as psychological. Also, the "psychological" dimension seems to be alluding to Big 5 traits. I'd suggest using Big 5 terminology here.

ANSWER: Thank you. We agree with the reviewer and replaced the “psychological” with “personality”. We corrected this throughout the revised manuscript.

p4. "LWIC uses pools..." LWIC typo; not sure what "pools" means in this context

ANSWER: We understand that the term “pool” would not be appropriate here and we changed with the term “set”. Sets in this context mean dictionaries of words (ie definite sets of words) identified as typical of specific psychological processes based on analyses of large amounts of texts (see ref 21).
p5. This isn't really a criticism, but I'm just interested in why so many of the summaries are written in Croatian (compared to say, French or German — languages that have a greater speaker population)?

ANSWER: We thank the reviewer for this question. Actually, translation of Cochrane information into Croatian is one of the main activities of Cochrane Croatia (see here: [https://community.cochrane.org/sites/default/files/uploads/inline-files/Croatia%20%20crowdsourcing.pdf](https://community.cochrane.org/sites/default/files/uploads/inline-files/Croatia%20%20crowdsourcing.pdf)), and we are currently preparing a publication on this (Behmen D, Marusic A, Puljak L. “Capacity building for knowledge translation: a survey about characteristics and motivation of volunteer translators of Cochrane plain language summaries”).

p6. "Scientific abstracts had significantly lower analytical tone that the other format". I think it would be helpful to spell out what "lower" and "higher" mean here.

ANSWER: This means that such texts used less words typical of analytical style (e.g. words and phrases implying causal connections, such as "therefore..." or "if...then..." etc.). This may be related to fact that scientific abstracts are highly structured. This is now clarified in the Results section of the revised manuscript.

p6. What was your motivation for analysing Big 5 traits? Also, I understood that these kind of tools (e.g. Watson Tone Analyzer) are intended to extract psychological variables from a single writer, and I suspect that most of these summaries are written by a committee. What impact do you think this has?

ANSWER: We understand that summaries have been written by different people with different personality characteristics and writing style. The idea behind it was to assess the resulting style in order to understand which traits are typically represented. Our aim was not to identify personality traits of the writer(s) but to understand whom the text may appeal to, as the psychological characteristics of the text have been show to affect the understanding of the text by a reader, information processing and decision making.

p7. It's not clear to me how the summary variables you use (analytical thinking, authenticity, etc.) map to LIWC categories as shown at this URL: [http://lit.eecs.umich.edu/~geoliwc/LIWC_Dictionary.htm](http://lit.eecs.umich.edu/~geoliwc/LIWC_Dictionary.htm)

ANSWER: The link above is to a sample dictionary that seems to be the older version. We used the LIWC2015 that is described in detail here:

p8. "The low readability of different textual formats for presenting finding of evidence synthesis in systematic reviews has important implications for the efforts in effective textual health evidence translation to different audiences." This is a bit difficult to understand

ANSWER: Thank you, this is now revised for clarity.

p8. "may adversely effect accuracy" what do you mean by "accuracy" here?

ANSWER: Thank you, this is now revised for clarity.

p8. "truthful information” I suggest not using the word "truthful" here.

ANSWER: Thank you, this is now revised for clarity.

p8. "the will like it and have confidence…” suggest reformulating

ANSWER: Thank you, this is now revised for clarity.

p9. "We used three different deep language processing tools" LIWC cannot be reasonably described as a deep language processing tool.

ANSWER: Thank you, we now use the term “language analysis tools”.

p9. "and the proposition to introduce more human, emotional element to enhance the understanding of basic health information” wording here is a bit strange.

ANSWER: Thank you, this is now revised for clarity.

p10. "The data sets used and/or analysed during the current study are available from the corresponding author on reasonable request" What does "reasonable request" mean in this context?
ANSWER: This statement is now rewritten to remove the requirement of a “reasonable request”. Our meaning for the “reasonable request” was that the data set could be used for research purposes.