Reviewer's report

Title: Automatically Identifying Social Isolation from Clinical Narratives for Patients with Prostate Cancer

Version: 1 Date: 13 Jan 2019

Reviewer: Jia-Wen Guo

Reviewer's report:

Background

It would be a better argument in this paper if there is an explanation for why we cared about to extract social isolation information from the EHRs. For example, what kinds of knowledge could be studied or generated by using the NLP to extract the social isolation information from narrative clinical notes.

The authors compared the outcome of the study with elderly population in the method and discussion sections. I assumed that it could be due to the prostate cancer patients tended to be elder males but this information was not mentioned in the background section.

Method

Under "Development of the lexicon for social isolation, this statement, "Social isolation is not commonly documented in the EHR including in clinical notes", was stated but the rationale of this statement is not clear.

Validated instruments were used to develop the social isolation lexicon. Only "Loneliness Scale" was listed. It would be clear to list all the instruments used to develop the lexicon.

Two domain experts generated five seed terms ("lack companionship", "feel left out", "isolated", "loneliness", and "lonely"). Comparing these seed terms to the Table 2, it seems like that "lack companionship", "feel left out", and "isolated" were not included in the "final" list; but there is no explanation about it. Is there any limitation by using the seed terms generated by behavioral science researchers?

Domain experts were used in the chart review (under Development of NLP algorithm to identify social isolation section") but it was not clear about who these two domain experts were and why they were qualified for the chart review. Were they the same experts to generate the seed terms?

The concept of social isolation is abstract. Without an adequate operational definition of social isolation for the domain experts could cause discrepancies during the manual chart reviews. However, the training for the domain experts to make sure they used the consistent operational definition of social isolation was not mentioned in the manuscript. Although the authors
mentioned the rater agreement is 97.4%; however, it is based on the "notes" but not based on the mentions. By the way, the percent agreement does not account for chance agreement.

"Two domain experts" were mentioned multiple times in the method sections; were they the same experts who are behavioral science researchers?

Results

Table 1 shows the note type in the training dataset. Was the distribution of the note types in the testing dataset similar to those in the training dataset?

Figure 1 was hard to read due to the very small font size and the truncated words.

In Table 2, the first word of some terms was capitalized. Do they have any meaning?

Based on the Table 3, this statement, "…more likely to be white and Medicare/Medicaid patients". Please also mentioned that the distribution of the race and insurance type of the "Positive" was similar to those presented in the "Negative".

In Table 4, it would be more informative if the note types of the false positive findings were provided.

Discussion

The authors mentioned "....the lexicon generated from the current study combines standard concepts and domain expert knowledge, our approach offers a more complete data extraction method". However, it was not very convinced. The domain experts only generated 5 terms and only 2 of them were listed in the final list. Moreover, what are those "standard concepts"?

Limitations

Limitations of this study were addressed properly.

The authors developed a NLP approach to extract the social isolation information but only 24 terms were included. It is possible that many terms used to describe the social isolation were not included since the definition or description of social isolation can be vary depending on the population.

Conclusion

Although the NLP algorithm in this study showed "a high accurate performance", it could be due to a limited number of the relevant terms used in the algorithm.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

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

Yes

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

Yes

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:

Acceptable

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