Reviewer’s report

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

Version: 0 Date: 01 Oct 2018

Reviewer: Sumithra Velupillai

Reviewer's report:

Overview:

This manuscript outlines a study on automatically identifying social isolation from clinical narratives for patients with prostate cancer using natural language processing methods. The study is performed on a large sample of electronic health records from the Medical University of South Carolina, including different clinical note types for a cohort of over 4,000 patients. The Linguamatics I2E version 5.3 software was used to develop the NLP solution, which included defining and developing a lexicon of terms for social isolation, as well as defining exclusion criteria to eliminate false positives. The data was split into a training/development set (75% of patients) and a test set, which was reviewed by a domain expert blinded to the query development.

Overall, this is a very interesting study which tackles an important but understudied problem in this field. The manuscript is in general well-written and structured, the study design is generally sound, the results are interesting, and the discussion is comprehensive with some enlightening examples.

There are, however, some concerns with this paper, that need clarifications and/or elaborations.

One major concern is the evaluation and manual chart review, particularly on the evaluation of 'control' notes. What was the motivation for choosing only 40 notes for this part of the evaluation? Given the obvious prevalence issue of this construct, this sounds like a very (too) small sample. Why was this evaluation not performed on all notes for a random sample of patients instead? That would more likely give a better estimate of potential false negative rates? Related to this, the approach to identify terms for the lexicon/terminology seems generally sound, but will inevitably not necessarily capture all potential synonymous variants. Did you consider other approaches to generate potential synonyms, e.g. using data-driven methods such as word embeddings (e.g. similar to Bejan et al.'s study (ref 10))? 

Further to this, as is mentioned in the manuscript, the identified concepts were more likely to be found in specific note types and sections - what was the prevalence of the different note types, and the distributions of found concepts in the different types? What was the motivation for including these particular note types in the study (e.g. radiation oncology)?
The background is concise and well motivated, but there is a lack of definitions and/or
discussions about the key concept (social isolation) - this would be important to include to better
understand the subsequent methodological work. For instance, the concept of 'social support' is
used sometimes as an important variable/concept which probably is synonymous to 'social
isolation', some more discussion on the nomenclature and definitions (e.g. from the
clinical/epidemiological perspective) would be important to include.

Other relevant studies that might be of interest, where identification of social risk factors using
NLP including social support have been included:

Greenwald JL, Cronin PR, Carballo V, Danaei G, Choy G. A Novel Model for Predicting
Rehospitalization Risk Incorporating Physical Function, Cognitive Status, and Psychosocial
Support Using Natural Language Processing. Med Care. 2017 Mar;55(3):261-266. doi:

Hospital Readmission and Social Risk Factors Identified from Physician Notes. Health Serv Res.

Some additional minor comments, questions and typos:

Data size: there are mismatching numbers provided about the training data set size: '55,516
clinical notes from 3,138 patients' (Abstract) vs. '3,138 patients (75%) as a training dataset with
150,990 notes to develop the lexicon and NLP pipelines' (Data source) - which number is
correct?

Chart review: in the section 'Development of NLP algorithm to identify social isolation' it is
stated that 'These chart review evaluations were done independently by a domain expert who was
blinded to the I2E query development', while in the section 'NLP algorithm performance' it is
stated that two domain experts performed the review. Please explain, and also clarify if this was
a blind double-review (if so, what was the agreement?) or if each chart was reviewed only by one
domain expert (if so, why?)

Table 1: very interesting with these examples and prevalence numbers. What was the motivation
for not conflating at least capitalized versions of the same concept to one frequency count (e.g.
'social isolation' and 'Social isolation')?

Resulting NLP approach: will any of this be made available to the research community (e.g.
resulting terminology, queries/rules)? What was the motivation for choosing this particular
(commercial) NLP system? Did you consider comparing results with other (open source)
systems?

Typos (might have missed some here, please ensure a thorough read-through and proof-read
before submission)*
Abstract: (HER --> (EHR)

Study setting: Intuitional --> Institutional

Development of the lexicon for social isolation: (Hughes et al., WILL SEND CITATION) --> remove

Development of the lexicon for social isolation: behavioral science research who --> behavioral science researcher who (?)

Development of the lexicon for social isolation: team came to consensus agreement --> team reached consensus agreement

Discussion: sematic --> semantic

To conclude, this manuscript presents a very interesting study of great relevance to the audience of this journal, but the methodological setup and study design needs to be clarified.

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.

No

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'

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/](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