Reviewer's report

Title: Increasing the Efficiency of Subject Identification: Automated Clinical Trial Eligibility Pre-Screening for Pediatric Oncology Patients

Version: 1  Date: 16 December 2014

Reviewer: Rachel L Richesson

Reviewer's report:

This manuscript describes an evaluation for matching algorithm in the context of patient screening for clinical trials. The algorithm that is evaluated (although not described) seems to be a promising solution to an important general problem, namely the need for automated methods to support the identification and possible matching of patients to trials.

The evaluation is well conceptualized and sensible. The results certainly demonstrate the role that automated matching systems such as this one evaluated here (and previously published in JAMIA in 2014 (citation #24). The discussion and conclusions are for the most part well-balanced and adequately supported by the data, and the limitations of the work are clearly stated.

However, the manuscript is not suitable for publication in its current form, and needs to be edited significantly, both for consistency and understandable writing style and for accuracy and clarity of content. This reviewer recommends that authors seek assistance to carefully revise the manuscript to address these two broad concerns. In addition to support from an editor proficient in scientific writing, one of the authors should closely look at this manuscript to make sure that all terms are used appropriately and accurately.

Major Compulsory Revisions

The question posed by the authors is important, but the description of the problem and its challenges could be better described in the background section. Specifically, the distinction between the two use cases of patient matching to trials and trial matching to patients is fundamental to the application of the algorithm and the design of this evaluation study, but it is only introduced in the results section, needs to be presented early on. More concerning, there are two sections of the document where the distinction between patient # trials versus trial # patients is not recognized, or is referred to incorrectly. (Specifically, see the sentences on page 3 line 59 “The rationale is to automatically rank the patients for a clinical trial (or vice versa)...” and p. 7 line 131 “We refer to this process as patient cohort identification…. Patient centered trial recommendation.”) Perhaps even the title should be reconsidered to clarify this – by either adding the word ‘matching’ or using terminology consistent with the rest of the paper. An authors that is familiar with clinical trials terminology and methodology should review the manuscript closely to make sure that the context is described accurately and completely.
Authors should more clearly describe and acknowledge any work upon which they are building. The introduction of previous work is insufficient. The first 3 sentences of page 4 are very confusing. Even if the work is previously described in ref 24 (JAMIA 2014) the algorithm should be described beyond “state of the art NLP and IE technologies.” What features are included? (e.g., keyword matching, negation, use of concepts from controlled terminologies or ontologies?) The algorithm description and the application description can be brief but authors should mention if and how this algorithm or approach for matching/ES is like or different from previous systems.

The sentence “Since a substantial proportion of meaningful information in EHRs is represented only in narrative text” (p.3, line 61) is not supported by a reference. Aren’t many important concepts (e.g., diagnosis, demographics, procedures, medications) in structured data? This is worth a mention, especially since the “state of the art NLP and IE technologies” are not described in this paper, it is not clear if you are using any structured data at all.

The 3rd sentence of page 4 (“The specific aim of our study is to … “) needs to be edited for accuracy and completeness. The notion of sensitivity is not mentioned in the abstract or results, and the use of patients and clinical data should probably be included in this first introductory sentence.

There are many instances where clinical trial related terms are not used accurately, and so an author needs to look at this. Examples are use of accrual on page 6, line 114. Line 118 mentions “Inter annotator agreement” which should be changed to inter-rater or screener. The whole discussion of Phase 1/2/3 trials on page 13 needs to be clarified.

Page 6, line 126-7: “….triggers manually derived from the eligibility criteria..” why not just say that patients were excluded by demographics?

The description of results is comprehensive, but needs editing. Much of this can be streamlined.

There are many tables and figures. Table 1 and perhaps Table 3 could be included in an online appendix perhaps. Table 3 is important but could be described and summarized in the manuscript narrative.

Figure 1 is presented as the sole description of eligibility criteria, but there is little description around it. It would be helpful to have a sentence or two in your background section to describe the complexity of eligibility criteria and the number and types of concepts they contain. It is possible that this figure could be dropped in lieu of a statement describing the nature and challenges of eligibility criteria.

Figure 2 is very nice, but is not described at all. It is presented in the manuscript as a “descriptive statistics of the algorithm” but important terms (e.g., ‘tokens’) are not defined. An overview of the main features of the system should be presented in the background or methods section, and then the results of Fig.2 will be presented in terms familiar to the reader.

Table 4 is a nice table (!) and should be described a little bit in the narrative
The sentence about usability and translational research in the Conclusion (p.15, line 291-2) is not supported by the study.

Minor Essential Revisions:

In the abstract and throughout, the phrase “eligibility criteria paragraphs” is used, and narrative eligibility criteria might be better term. Also in the abstract under methods section, the sentence “The algorithm then executed …its function…” could be better worded. Abstract conclusion section: reconsider the words ‘involve’ and ‘currently’; you might try ‘enable’ and ‘often’.

Background section page 2 first sentence: Consider rewording this to something like: “Although several reports have described positive experiences leveraging EHR information to facilitate trial recruitment, ES is still conducted manually in most cases.”

Sentences beginning on lines 54 and 55 on page 2 need some editing. Can you support the statement that it is insufficient staff or labor intensive screening specifically that are responsible for ‘insufficient enrollment’? If not, you might reference those more broadly as “challenges” for trial recruitment and present it more anecdotally.

You could possibly delete the two sentences include in p.3 lines 57 and 58. If your data source is clinical notes, then I would refer to that directly, rather than say “..patient EHRs (e.g., clinical notes).

The sentence (p.3 line 67) that mentions “lack of access of real-world EHRs” needs to be edited for accuracy and appropriateness for this audience. Any reader from an organization with an HER that is looking at this work will indeed have a “real world EHR.” Not sure what access you are referring to.

Page 5, line 111 – instead of “eligibility screening”, should you say “enrollment”?

The discussion uses the term “demography based” and I think “demographics-based” would be a better term. And somewhere you should mention which concepts your algorithm looks at (e.g. gender, age, race (?) ).

Not sure why you would use the term ‘rudimentary’ to describe your application. The discussion and conclusions are well balanced and adequately supported by the data, and the limitations of the work are clearly stated.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Not suitable for publication unless extensively edited

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.