Reviewer's report

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

Version: 1 Date: 28 August 2014

Reviewer: Philip Payne

Reviewer's report:

This paper describes the design and evaluation of an NLP/IE-based pipeline intended to facilitate the pre-screening of pediatric patients for oncology-related clinical trials. The approach as described shows promise, leveraging a process of computational modeling and integrative analysis of both eligibility criteria and clinical phenotype data so as to reduce end-user workload relative to the identification of patients with a high likelihood of being eligible for a given trial. The manuscript is concise, well written, supported by relevant background and prior work in the field, and makes a notable contribution to the state of Clinical Research Informatics knowledge and practice. This paper should be of great interest to readers of BMC Medical Informatics.

Overall, this reviewer only has three major suggestions for potential additions/modifications to the paper that may serve to enhance its overall impact, as follows:

Compulsory Revisions:

The approach used by the investigators employs only a small amount of structured data extracted from the EHR (largely related to demographics), and heavily relies upon the extraction of clinical phenotype data from unstructured narrative text. However, prior work in this area has shown numerous areas of domain coverage in which eligibility criteria can be satisfied using structured data as found in the EHR (for example, data elements related to prior history, treatment, laboratory values, etc.). The rationale for why such data was not used in this study, and even between, and comparative analysis of the performance of discrete data versus NLP-derived data versus a combined approach, would serve to greatly enhance the generalizability of this work.

Furthermore, much of the prior work in this field has described challenges related to the accurate and comprehensive modeling of eligibility criteria in a computational tractable manner. Issues described in such reports have included the ability to address temporal constraints/reasoning, the granularity of formal concept mapping relative to the sub-language commonly encountered in such criteria; and the prevalence of complex/nested logic relative to qualifying statements in those criteria. The current study as described in this report is effectively silent on such issues. Did the authors encounter them when mapping criteria to data? How did they address them? Did they use alternative modeling
approaches that overcame such barriers? These questions should be addressed more fully in the manuscript.

Finally, the evaluation of workflow implications for the purposes of supporting the utility of the described approach appears to use a combination of retrospective or simulated data. While the resulting assertions may be accurate, their validity would have been greatly enhanced by a prospective study of the described pipeline in real-world clinical research settings. This limitation would at a minimum be cited in the report, and ideally, further data to such an effect would be collected, analyzed, and reported upon so as to empirically defend the papers conclusions.

Minor Essential Revisions:
In addition to the preceding major comments, the paper would also benefit from the inclusion of more comprehensive and descriptive captions for all figures and tables, as well as a more explicit enumeration of future directions for the work (especially for those activities that would address noted limitations).

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

**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