Improving Clinical Trial Cohort Definition Criteria and Enrollment with Distributional Semantic Matching

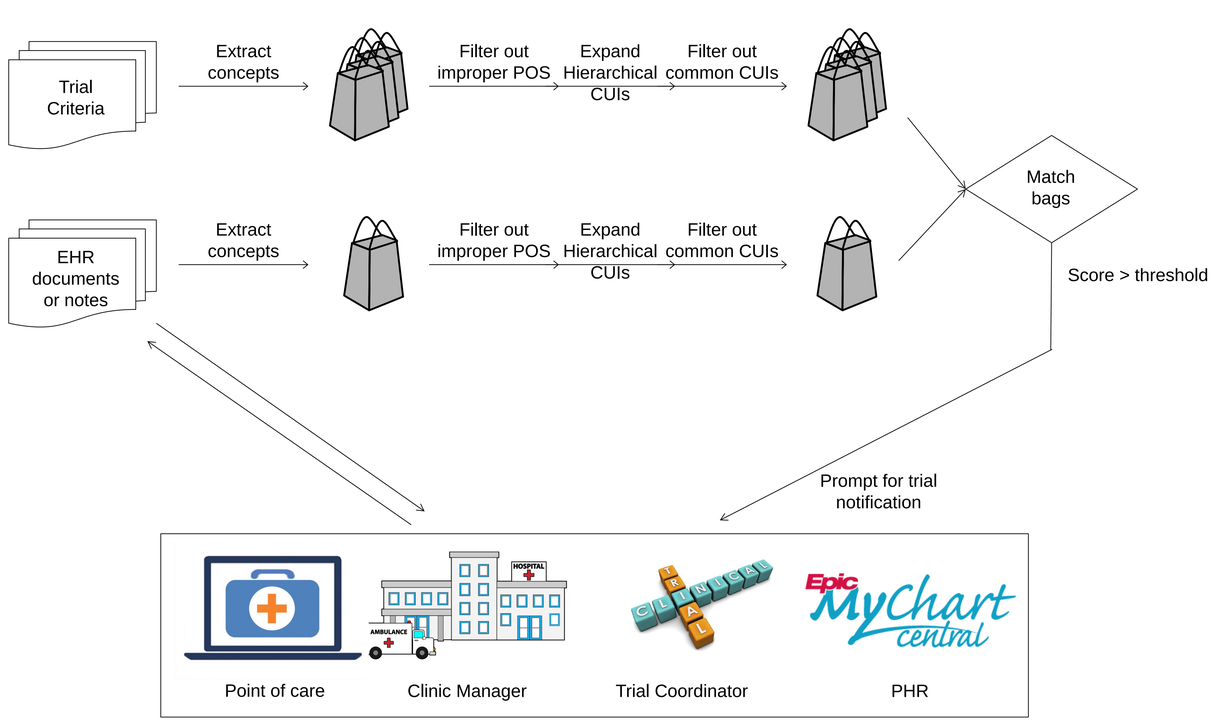
Jianyin Shao, PhD[1], Ramkiran Gouripeddi, MBBS, MS[1,2], Julio C. Facelli, PhD[1,2]

[1]Department of Biomedical Informatics, [2]Center for Clinical and Translational Science, University of Utah, Salt Lake City, Utah, USA

Introduction

- Evidence-base medicine relies on well-designed and performed reproducible research.
- Clinical trials are the gold standard of experimental design for examining effects of clinical intervention on patients or populations.
- Current clinical trial cohort recruitment approaches may not promote reproducible research due to
  - Ambiguous cohort definition and varied interpretation by clinical trial coordinators.
  - Biases of selection of cases and controls.
- Formal semantic matching patient medical record with appropriate trial eligibility criteria could improve the reproducibility of the clinical trials.

Flow of Information

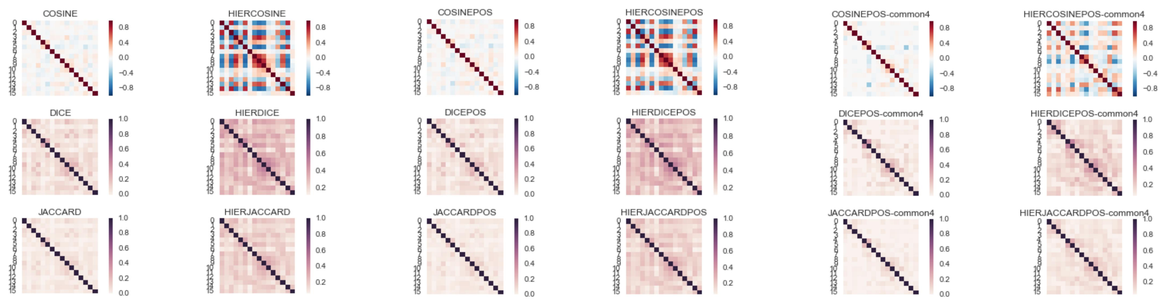


Method

- Concept Bag: clinical trial eligibility criteria and/or patient medical records can be represented by a set of UMLS concepts.
- Hierarchical Concept Bag: trial criteria and medical records can be represented by a set of UMLS concepts and their ancestors.
- Concepts extracted by MetaMap.
- Similarity metrics: Jaccard, Dice, Cosine.

Results

Pair-wise similarity for 4 sets of 4 trial eligibility criteria using different pre-processing and metrics



All CUIs | Retention of CUIs of specific POS* | Removal of common CUIs**

* Retained noun, verb, adjective, adverb and numerical cardinal CUIs.
** A common CUI list of 1648 CUIs was compiled by manual review of 195 randomly selected trial eligibility criteria (all trials ending with “000”)

Discussion

- On the 4x4 curated criteria sets, HCB with Dice yields the best performance with relatively high intra-set similarity and relatively low inter-set similarity.
- HCB can provide similarities that match human expert cognition.
- Pre-processing CUIs improves the similarity comparison, especially filtering out CUIs that are too common or general.
- Other approaches can be used to refine concepts, such as using UMLS concept semantic types.
- For matching patient records, metadata of the EHR can be used to select appropriate semantic types of the CUIs for similarity comparison.

References

1. Bradshaw RL. Concept Bag: A New Method for Computing Similarity. Ph. D. Dissertation, University of Utah; 2015.
2. Aronson a R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;17–21.
3. Wilcox A, Natarajan K, Weng C. Using Personal Health Records for Automated Clinical Trials Recruitment: the ePaIRing Model. Summit Transl Bioinforma. 2009;2009:136–40.

Acknowledgements

This work has been supported in part by the National Library of Medicine Grant# T15LM007124 and National Center for Advancing Translational Sciences of the National Institutes of Health Award# UL1TR001067.

Contact Information

Jianyin Shao
jianyin.shao@utah.edu

The University of Utah
Biomedical Informatics