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

- 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.

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

<table>
<thead>
<tr>
<th>All CUIs</th>
<th>Retention of CUIs of specific POS*</th>
<th>Removal of common CUIs**</th>
</tr>
</thead>
</table>

* 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


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