Towards Lexical, Semantic and Similarity Search in Phenotype Libraries

Ramya Tekumalla, Raj Manickam, Yen Low

Background

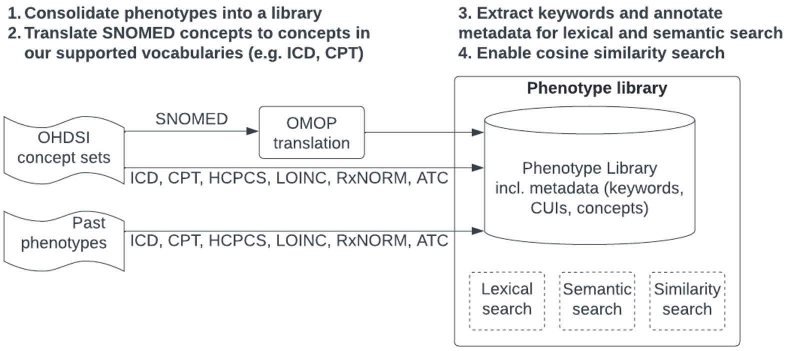
Numerous phenotype libraries[1] have emerged to enable more transparent and reproducible observational studies. Many are community efforts to crowdsource phenotype definitions, some of which have been curated and empirically validated. However, this has also caused much lexical and semantic variability, depending on how the phenotypes were named and defined in the vocabularies of interest. Such variation has made phenotype search and selection challenging.

Our objective is to improve search by utilizing phenotype similarity, based on the phenotype's concept sets, particularly administrative codes like ICD, CPT, HCPCS, LOINC, RxNORM and ATC which are used widely for electronic health records and claims research. Using such widespread codes as features for our search space has several advantages: they relate phenotypes (e.g. HIV, kaposi's sarcoma) that may not be obvious from lexical or semantic search, and they provide an interoperable search space from which other vocabularies (e.g. SNOMED) may be translated into.

Here we explain how we 1) constructed a phenotype library from readily available definitions like OHDSI's concept sets[2] and definitions used by clinical informaticians on our Atropos Health platform, 2) translated SNOMED definitions into administrative codes supported by our platform, 3) enriched phenotypic metadata with text descriptions and their UMLS entities for improved lexical and semantic search, and 4) performed similarity search which resulted in overall more phenotype choices.

Methods

Figure 1. Methods overview


1. Consolidate phenotypes into a library
2. Translate SNOMED concepts to concepts in our supported vocabularies (e.g. ICD, CPT)
3. Extract keywords and annotate metadata for lexical and semantic search
4. Enable cosine similarity search
Phenotype library
OHDSI concept sets
SNOMED
OMOP translation
ICD, CPT, HCPCS, LOINC, RxNORM, ATC
Phenotype Library incl. metadata (keywords, CUIs, concepts)
Past phenotypes
ICD, CPT, HCPCS, LOINC, RxNORM, ATC
Lexical search
Semantic search
Similarity search

Data collection

At this initial stage, we focused on the simplest forms of phenotypes which are formulated as concept sets[1], leaving out those composed from more complex rules. We collected 703 concept sets from the OHDSI phenotype library[2] and another 1,832 definitions by Atropos Health clinical informaticians in the period July 2021 to May 2022. The Atropos Health definitions were written in our temporal query language[3] (e.g. var X = ANY(ICD10="F10.1", ICD10="F10.2", ICD10="F10.9")) whose codes (e.g. F10.1) were extracted using regular expression rules. We limited OHDSI concept sets to those defined by vocabularies supported in the Atropos Health platform (ICD, CPT, HCPCS, LOINC, RxNORM, ATC). As

OHDSI concept sets are published in standard vocabularies, we translated concepts defined in SNOMED into our supported vocabularies using the vocabulary mappings in OMOP CDM 5.4[4]. While their definitions may shift during translation, such translation allows us to expand our collection for practical reuse while still providing reasonable definitions for further empirical evaluation and expert review.

To address descendants and exclusions in the concept sets, we first retrieved all concepts marked for inclusion (e.g. malignant neoplasms) and their descendants unless otherwise mentioned. We then removed those marked for exclusion (e.g. nonmelanoma skin cancer) such that we obtained a net set of concepts describing the desired concept set (e.g. malignant neoplasms excluding nonmelanoma skin cancer). Exclusions that are not subsets of inclusions will be separately handled in future work when the library may support more complex compositions of rules as phenotypes.

Lexical search

We enriched the phenotypic metadata by utilizing its text description, SNOMED labels if any in addition to its name so all will be used for text matching.

Semantic search

Text descriptions (e.g. surgical bowel resection) related to the phenotypes were annotated using ScispaCy[5] loaded with the *en_core_sci_lg* model and the UMLS entity linker. The highest scoring concept unique identifier (CUI) and its canonical name (e.g. C0741614, intestine resection) were additionally saved as metadata. During semantic search, users could look up "bowel resection" which has been annotated as C0741614 and find all other concept sets also annotated by the same CUI (e.g. intestine resection).

Similarity search

To create the feature space for similarity search, we extracted the name and concepts in each concept set, treating each concept set as a document of a bag of concepts. These features were further transformed to enrichment ratios using scikit-learn's TF-IDF vectorizer. Two documents (i.e. concept sets) were similar if their pairwise cosine similarity exceeded 0.7. At this initial phase, we limited this work to only documents (i.e. concept sets) composed of only ICD-10-CM codes.

Results

Table 1 presents several concept sets returned via lexical, semantic and similarity search. As expected, each search mode returned slightly different concept sets depending on the basis of search.

Semantic search generated concept sets (e.g. inflammatory bowel disease, irritable bowel syndrome) whose names did not contain the keywords used for lexical search (e.g. IBD). Sometimes, semantic search returned a broader concept set (e.g. eye extraintestinal manifestations) of which the keyword condition (e.g. uveitis) is part of.

Similarity search was effective at returning related conditions. For example, several diabetic complications and psoriatic complications showed up in similarity search but not lexical or semantic search when searching for the root condition "type 2 diabetes mellitus" or "psoriatic arthritis" respectively.

In such a case, if one expects a specific definition, it may seem that similarity search would return many false positives. But if one prefers more suggestions for review and modification, searching via multiple modes provides additional choices. The quantity and quality of the choices can be further improved by enabling/disabling certain search modes, supporting lexical variants by text normalization, supporting semantic similarity using appropriate ontologies[6], tuning similarity thresholds, and ranking the concept sets by appropriate metrics.

Table1. Example concept sets found with different search methods

Keyword for search	Concept sets with lexical search	Concept sets with semantic search	Concept sets with similarity search
uveitis	4 - {uveitis, uveitis guo, uveitis, anterior uveitis in juvenile idiopathic arthritis}	4 - {uveitis, uveitis guo, uveitis, eye extraintestinal manifestations}	3 - {uveitis, uveitis guo, uveitis}
malignant neoplasm excluding non melanoma skin cancer	1 - {malignant neoplasm excluding non melanoma skin cancer}	1 - {malignant neoplasm excluding non melanoma skin cancer}	3 - {malignant neoplasm excluding non melanoma skin cancer, malignant neoplastic disease, lymphoma}
type 2 diabetes mellitus	2- { type 2 diabetes mellitus, type 2 diabetes mellitus diagnosis}	2 - {type 2 diabetes mellitus, type 2 diabetes mellitus diagnosis}	5 - {type 2 diabetes mellitus, retinopathy due to diabetes mellitus, diabetic ketoacidosis, type 2 diabetes mellitus diagnosis, miller codes for diabetes mellitus}
hepatitis c	5 - {hepatitis c, viral hepatitis c, viral hepatitis c, viral hepatitis c, hepatitis c}	5 - {hepatitis c, viral hepatitis c, viral hepatitis c, viral hepatitis c, hepatitis c}	5 - {hepatitis c, viral hepatitis c, viral hepatitis c, viral hepatitis c, hepatitis c}
ibd	1- {ibd}	4 - {ibd, penetrating ibd, irritable bowel syndrome, inflammatory bowel disease}	6 - {ibd, inflammatory bowel disease, ulcerative colitis, ulcerative colitis, ulcerative colitis, ulcerative colitis}
psoriatic arthritis	3 - {psoriatic arthritis exclude dactylitis mutilans , arthropathy component exclude psoriatic dactylitis and arthritis mutilans, psoriatic arthritis or arthropathy}	3 - {psoriatic arthritis, psoriatic arthritis exclude dactylitis mutilans, psoriatic arthritis or arthropathy}	10 - {psoriatic arthritis, psoriasis, psoriasis, psoriasis component, psoriatic arthritis exclude dactylitis mutilans, psoriasis excluding guttate psoriasis and palmoplantar pustulosis, psoriasis, psoriasis with arthropathy, psoriasis, pustular psoriasis}

Conclusion

This study demonstrates how we enabled various search modes in a phenotype library. First, we extracted keywords, UMLS entities and CUI from text descriptions as metadata for improved lexical and semantic search. Second, we enabled cosine similarity search by treating each concept set as a document of a bag of codes (and transformed to TF-IDF enrichment ratios). As each search mode returned slightly different concept sets, taken as a whole, they recommend overall more concept set choices for expert review and modification.

Ongoing work is directed towards improving search in several ways, e.g. allow users to select search mode(s), perform text normalization to handle lexical variants, support semantic similarity search using appropriate ontologies[6], better feature representation using pre-trained embeddings, tuning similarity thresholds, and ranking the concept sets by appropriate metrics. Other library improvements include sourcing more concept sets, supporting more vocabularies, and supporting more complex formulations of phenotypes beyond concept sets.

References/Citations

1. Chapman M, Mumtaz S, Rasmussen LV et al. Desiderata for the development of next-generation electronic health record phenotype libraries. GigaScience. 2021;10
2. OHDSI PhenotypeLibrary. Available from: https://github.com/OHDSI/PhenotypeLibrary (retrieved 22-Apr-2022).
3. Callahan A, Polony V, Posada JD, Banda JM, Gombar S, Shah NH. ACE: the Advanced Cohort Engine for searching longitudinal patient records. Journal of the American Medical Informatics Association. 2021;28:1468-1479.
4. OHDSI. Common Data Model. Available from: https://github.com/OHDSI/CommonDataModel (retrieved 04-Feb-2022).
5. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task. 2019
6. Gkoutos GV, Schofield PN, Hoehndorf R. The anatomy of phenotype ontologies: principles, properties and applications. Briefings in Bioinformatics. 2018;19:1008-1021.