Evaluating the OHDSI Phenotype library concept sets using Large Language Models.

Dmytro Dymshyts MD[1,2], Joel Swerdel, PhD MS MPH[1,2], Anna Ostropolets MD PhD[1,2], Azza Shoaibi, PhD[1,2], Patrick Ryan, PhD[1,2], and Martijn Schuemie PhD[1,2]

[1]Observational Health Data Analytics, Global Epidemiology, Janssen Research and Development, Titusville, NJ, USA

[2]Observational Health Data Sciences and Informatics (OHDSI), New York, NY, USA

Background

The Observational Health Data Sciences and Informatics (OHDSI) community has developed a publicly accessible, version-controlled Phenotype Library (OHDSI PL) to guide real-world evidence towards the FAIR principles: Findability, Accessibility, Interoperability, and Reproducibility[1]. This library aims to foster the submission and retrieval of high-quality cohort definitions, cataloging of metadata, attribution and promotion of discovery and reuse in scientific research.

When developing phenotype definitions for health conditions, one of the first steps is to develop the list of codes used to determine the phenotype. In OHDSI, source concepts from various terminologies map to standard concepts defined by the OHDSI Standardized vocabularies [2].

To check which medical events are captured by our phenotype definitions, we need to investigate the source codes our definitions result into.

Objective

The objective of this study is to utilize the use Large Language Models (LLMs) to check if source codes obtained from concept sets in OHDSI PL correspond to the intended idea of the concept set.

Methods

We ran the LLM process through 584 cohorts, containing 1828 concept sets in total.

To evaluate the impact of the detected source codes , we looked at total record counts in various data sources:

US commercial claims:Health Verity Comprehensive Claims and EHR - Closed Claims Enrollment

- Merative(TM) MarketScan(R) Commercial Claims and Encounters Database
- Merative(TM) MarketScan(R) Multi-State Medicaid Database
- Merative(TM) MarketScan(R) Medicare Supplemental and Coordination of Benefits Database
- Optum’s Clinformatics(R) Extended Data Mart, Date of Death (DOD)

US EHR data:

- Optum EHR

EHR data from France, Germany, Australia and Japan claims data (JMDC).

Each concept set then was resolved into the list of source codes. In this work we focused on the US source medical vocabularies, such as 'ICD10', 'ICD10CM', 'CPT4', 'HCPCS', 'ICD9CM', 'ICD9Proc', 'ICD10PCS', 'LOINC', 'NDC'. To focus only at the most important codes we only kept the source concepts with a record count of 2 or more percent of the record count of all concepts in the concept

set. Note: the record count is the sum of the number of occurrences of a given concept in all databases in our network. This resulted in 11855 pairs of "concept set name – source concept" .

For this study we used OpenAI Model GPT-4o, trained through October, 2023. Procedural calls to the application programming interface (API) for the LLM were made using the R platform.

The LLM was prompted to evaluate each pair of concept set/source concept. The prompt used (**bold** is a text of prompt itself, *italic* – comments:

System prompt: "**You are a medical terminologist**" – *this allowed to focus on the specific area*

Prompt:

concept set name: %s - *concept set name as defined in cohort definition*

term: %s - string containing *vocabulary_id, concept_code and concept name – the ICD9 and ICD10 codes are often used in the literature, thus the codes might be important for the model*

Check if medical term belongs to the concept set
Output format:
 # Reasoning:
- Reasons why the term should belong to the concept set – TRUE
- Reasons why the term should not belong to the concept set – FALSE
 # Final answer: [TRUE or FALSE] (TRUE if term belongs to concept set, FALSE if term doesn't)

Then the output was parsed, and we got the binary answer: TRUE if concept belongs to the concept set, FALSE if it doesn't belong according to the LLM.

LLM output was validated by clinician by reviewing random pairs of "concept set – source concept".

Results

912 concepts in 200 cohorts were identified as potentially not belonging to the assigned concept set. The median proportion of record counts of concepts marked as FALSE was 12.3%.

Results of clinical review:

A random set of 100 TRUE and 100 FALSE concept set – source concept pairs were reviewed by a clinician. It was confirmed that All reviewed LLM TRUE cases stood for cases where the concept belonged to the concept set. 57 of LLM FALSE cases stand for the cases when the concept shouldn't belong to a concept set as determined by clinician, other 43 cases need more context, as It's not always clear what is meant by the concept set based on its name. In the future we are going to investigate whether source concepts flagged as FALSE originate from the initial cohort definition design or appear due to mapping peculiarities.

Please see the examples below:

The LLM determined the Lymphocytopenia concept as FALSE in the "Neutropenia or Agranulocytosis" concept set with the following reasoning:

 # Reasoning:
- Reasons why the term should belong to the concept set - FALSE
- The concept set "Neutropenia or Agranulocytosis" refers to conditions involving a decrease in

neutrophils or granulocytes, which are specific types of white blood cells.
- The term "Lymphocytopenia" (concept_code: D72.810) refers to a decrease in lymphocytes, ***which are a different type of white blood cell.***
- Lymphocytopenia is not directly related to neutropenia or agranulocytosis, as it involves a different subset of white blood cells.
- Reasons why the term should not belong to the concept set - TRUE
- Lymphocytopenia is distinct from neutropenia or agranulocytosis in terms of the type of white blood cells affected.
- The concept set specifically focuses on neutrophils and granulocytes, not lymphocytes.
 # Final answer: FALSE

Outstanding cases are the combination drugs where one of the ingredients was indicated for the disease of interest, while in combination with other they are indicated for a different disorder:

"HIV drugs" concept set including "{20 (nirmatrelvir 150 MG Oral Tablet) / 10 (ritonavir 100 MG Oral Tablet) } Pack [Paxlovid 5-Day]" = FALSE
the Reasoning:
- ***Paxlovid is an antiviral medication used for the treatment of COVID-19, not HIV. It contains nirmatrelvir and ritonavir, where ritonavir is a protease inhibitor commonly used in HIV treatment to boost the effectiveness of other HIV drugs. However, in Paxlovid, ritonavir is used to boost nirmatrelvir's effectiveness against SARS-CoV-2, not HIV.***
- The concept set "HIV drugs" specifically refers to medications used for the treatment or management of HIV infection. While ritonavir is used in HIV treatment, the combination of nirmatrelvir and ritonavir in Paxlovid is not intended for HIV treatment.
- Reasons why the term should not belong to the concept set:
- Paxlovid is not indicated for the treatment of HIV; it is specifically approved for the treatment of COVID-19.
- The presence of ritonavir in Paxlovid does not make it an HIV drug, as its purpose in this combination is unrelated to HIV treatment.
Final answer: FALSE

Another example: "Antiarrhythmics" concept set including "20 ML epinephrine 0.005 MG/ML / lidocaine hydrochloride 20 MG/ML Injection [Xylocaine with Epinephrine]" = FALSE.

In this case Epinephrine was a part of combination meaning that it was not an antiarrhythmic but a local anesthetic. These drug combinations were usually overlooked by the manual review.

In some cases, LLM got confused, for example:

"Uveitis Anterior and intermediate uveitis" concept set includes "Posterior cyclitis" = FALSE, the reasoning includes the argument that concept set has "anterior" and "intermediate", and the concept has "posterior" in its name. While the posterior cyclitis is the intermediate uveitis according to the anatomy of the eye.

Source concepts marked as FALSE by the LLM but can't be assessed as we don't know the full intent of the concept set, for example: Myalgia including the "Atherosclerosis of native arteries of the extremities with intermittent claudication". We don't know if Myalgia includes all the underlying causes or not.

Conclusion

The usage of LLM provides a helpful check on the included source concepts in cohort definitions, especially in the cases hidden from human review (combination drugs with different indications which sit deep in the hierarchy), cases when humans can mix up close terms (Neutropenia and Lymphocytopenia – both are a decrease of white blood cells). Also, this work highlights an importance of proper description of concept set names – it's important both in the manual and in LLM-driven review of concept sets. The future work includes contacting the cohort authors and maintainers to check the existing definitions as well as to improve the future process so it will include more comprehensive descriptions.

References

1. Wilkinson, M.D., et al., The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 2016. 3(1): p. 160018.
2. Reich, Christian, et al. "OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization." Journal of the American Medical Informatics Association, vol. 31, no. 3, Mar. 2024, pp. 583–590, https://doi.org/10.1093/jamia/ocad247.