Task 2: Aligning indications from multiple data sources

Code is not required for this section, but if you do write any code to explore your ideas please include it in your submission.
We want to present all of the data we have about each indication on an Indication Report page. We have data relevant to indications from 3 different public sources (ex: ClinicalTrials.gov, ChEMBL, pharos.nih.gov) but each source uses a different biomedical ontology to identify indications. One uses ICD10 codes, another uses MeSH ids, and another uses MONDO ids. The goal is to create a deduplicated list of indications so we can link to all of the data from these public sources for a given indication. We know that each source contains some indications not present in any of the others, some that are present in only 2 of the sources, and some that are present in all 3. Incorrectly creating the deduplicated list of indications could either result in the same indication being present multiple times, or indications that are not actually the same being linked together.

5. What tools or datasets already exist that could be relevant for this task?

Provide a brief summary of each, and your initial thoughts on the strengths and weaknesses of each. Call out any additional logic you would expect to have to implement on top of the existing tools to create a full working implementation.

a. **UMLS (Unified Medical Language System):** Developed by National Library of Medicine, UMLS integrates and distributes key terminology, classification and coding standards. It is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems. It can map between various biomedical ontologies, so it could be useful for converting ICD10 codes, MeSH ids, and MONDO ids into a common form.

Strength: Comprehensive coverage of biomedical terms in Metathesaurus (https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_001.html) with unique hierarchical identifiers (https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html).
Weakness: UMLS is only licensed to individuals and not groups or organizations (https://www.nlm.nih.gov/research/umls/index.html).

b. **BioPortal**: Developed by the National Center for Biomedical Ontology, it is a comprehensive repository of biomedical ontologies and offers searching, sharing, visualizing, and analyzing biomedical ontologies, terminologies, and ontology-based annotations. It has >95M mappings with >1K ontologies including ICD10, MeSH, and MONDO

Strength: It provides mappings between two or more terms in different ontologies both in the browser and programmatically (https://www.bioontology.org/wiki/BioPortal_Help#Mappings_Tab)
Weakness: Certain OBO-defined ontologies may be incomplete (https://www.bioontology.org/wiki/BioPortal_FAQ#incomplete_is-a_paths)

c. Matching Evaluation Toolkit (MELT) is a powerful framework for developing, evaluating, and packaging ontology matching systems. It is optimized to be used in Ontology Alignment Evaluation Initiative (OAEI), a coordinated international effort to organize ontology matching systems.

Strength: Supports multiple packaging and evaluation formats (https://dwslab.github.io/melt/matcher-packaging)
Weakness: Some MELT libraries do not work when evaluating with the SEALS client due to inconsistencies in dependencies (https://link.springer.com/chapter/10.1007/978-3-030-33220-4_17#Sec4)

Additional steps to create a working implementation:

(i) Disambiguation Logic: In case there are conflicting mappings in the source, proper disambiguation logic needs to be set up.
(ii) Confidence Scoring: If a tool provides multiple matches or mapping choices, additional scoring logic would be required to either select the best match or display multiple choices.

6. Write pseudocode for one of the approaches considered in Question 5.

We could use the Concept Unique Identifiers (CUI) in UMLS to approach this problem:

```
Initialize an empty dictionary for deduplicated indications

for each source in [ClinicalTrials.gov, ChEMBL, pharos.nih.gov]:
    for each indication in source:
        map indication to UMLS CUI using the relevant UMLS mapping
        if UMLS CUI is already in dictionary:
            append the indication to the existing list
        else:
            create a new list with the indication and add to dictionary

The deduplicated list of indications is now the keys of the dictionary
```

7. How would you evaluate the performance of your algorithm?

The Ontology Alignment Evaluation Initiative (OAEI) uses precision, recall, and F-1 scores to assess performance (https://ceur-ws.org/Vol-3063/oaei21_paper0.pdf). Another option is to compute the Mean Absolute Deviation (MAD) of the Jaccard similarity index as shown in the MELT paper (https://link.springer.com/chapter/10.1007/978-3-030-33220-4_17#Sec11)

8. What technology stack would you use for any data storage, processing, data access, analysis, model generation, or visualization you think is relevant to this task?

a. I think we should be able to use simple python code for the first step of pulling data from ClinicalTrials.gov that has a REST API, ChEMBL that offers a Web Services python client, and Pharos that appears to have a sophisticated GraphQL based API. We could also consider SQL.

b. Since we're dealing with structured relational data (indications, their sources, and their codes in different schemas), we could consider using a relational database like PostgreSQL for data storage.

c. For data processing and analysis, we could consider using SQLAlchemy that offers python-based Object Relational Mapping capabilities.

d. We could consider constructing Knowledge Graphs to learn ontology mappings.

e. For visualization on the Indications Report page, we could use Tableau, or convert it into a web app, by using Flask and deploying it on Google Cloud or AWS.