Medication Augmentation

Goal: To expand the conmeds.yml so that it comprehensively captures all generic and brand names for each drug class.

Our **normal workflow**:

1. Clinical scientists determine the most relevant drug classes to be included for a specific indication. List a variable named as `taking_{drug-class}` in the assemble schema/data specs google sheet https://docs.google.com/spreadsheets/d/1GObmXziHofjEK6EEMaI5_axzAx6jIsEBJ7NgO-4Ha80/edit?usp=sharing. The description will include notable medications for examples (which can be added with more examples if needed with a correct name).
    a. The variable column should stay untouched.
    b. The description column could be augmented. And new data specs should be returned if the description column is modified.

[Ideally, steps below to be fully automated].

2. A base `conmeds_defaults.yml` is created for the indication containing relevant keys of `taking_{drug-class}` with value of an array of medication names. Plombier will parse the values in the array using regex (I think this part of the pipeline is here https://github.com/unlearnai/core-libraries/blob/e4aac20d72eb9d6e814482ec80d3e78d60ab1527/packages/plombier/src/plombier/core/domains/conmeds/meds_enrich.py#L158).
3. Data scientists review the raw data and find the medication name column. If the medication name exists in the description columns of `taking_{drug-class}`, then a key named `taking_{drug-class}` will be added to the conmeds.yml file with the generic and brand names mentioned in the description.
    a. However, the names included in the description are not exhaustive.
    b. Often, data scientists need to take a closer review at the medication name in the raw data and googling if any of them should be added in.
4. The unique values of the medication column in the raw data is used to add to the array of medications in the yaml.
    a. Depending on the list length or complexity of the unique values in the source data, we will manually sort or use an LLM to sort into the corresponding arrays. Using an LLM has a fair amount of false negatives, which we hope to improve using web scraping to better contextualize the result.
5. Note: the conmeds.yml file can include typos to capture as many relevant medications into a `taking_{drug-class}` column.

Some **evaluation metrics/intermediate steps** to look out for:

1. What is the column in the raw data file that the claude agent used to extract medication names? Because there could be more than one drug name-like columns sometimes (see raw data example 2 below).

2. How many generic and brand names did Claude agent add on top of the examples that already exist in the descriptions for the ‘taking_{drug class}”? What is the improved percentage of medication matching before and after?

Links to assets:
Data specs google sheet:
https://docs.google.com/spreadsheets/d/1GObmXziHofjEK6EEMaI5_axzAx6jIsEBJ7NgO-4Ha80/edit?usp=sharing
Raw data example 1: (where there is just one drug name column)

- s3://unlearnai-prod-data-acquisition-team/20250728_msk_chord_2024/data_timeline_treatment.txt (also copied this to Gdrive folder https://drive.google.com/file/d/1VTv-C24nuBdzX-th5Y7e3ys5HHTDdRwa/view?usp=drive_link for easy access)
- The drug name column is ‘AGENT’.

Raw data example 2: (where there are multiple drug name-like columns)

- Link to data https://drive.google.com/file/d/1Ok1I86EFvbvfz_KYRZNE5XYjpxNDkPPH/view?usp=drive_link in Gdrive folder.
- The drug name column is ‘DRUGDTXT’. While the column ‘ACTTRTXT’ looks like a drug name column, it’s actually the column for treatment assignment during the trial and should not be used for medication mapping.