**Counting Disease Terms in Disease Node**

This study aims to identify and quantify the various types of data within the Disease node of the Condensed Knowledge Graph. The Disease node contains a total of 358,197 records, although not all are classified as diseases. To categorize these records, semantic types are extracted from the Unified Medical Language System (UMLS). A Python script is used to fetch the semantic types for each record, processing the data in batches due to its large size. The unique semantic types and their respective counts are detailed in the file "Unique\_Types\_Count.xlsx." Note that the counts are approximate, with a potential variation of ±10. The count for the semantic type 'disease or syndrome' in the Disease node, as determined by this study, is 40,554. All records in the Disease node are sourced from the SNOMED CT US vocabulary.

* A small set of input data given to the python code get\_semantic\_types\_for\_a\_list\_of\_strings.py is in strings.txt file.
* The output data from the python code is stored in the output.txt file.
* The cypher queries used to get the input data are listed in Cypher\_queries.txt file.
* The complete set of output data is stored in Results.txt file, and the excel sheet is in Results.xlsm file
* The python code used to convert the results from txt format to csv format is given in Txt\_to\_CSV.ipynb file.
* To run the python code get\_semantic\_types\_for\_a\_list\_of\_strings.py, input file name, output file name and the API\_KEY from the UMLS License Profile must be provided.
* Run command is : python3 get\_semantic\_types\_for\_a\_list\_of\_strings.py -k YOUR\_API\_KEY -i strings.txt -o output.txt

**Getting Hierarchy for Each Disease Term**

To find the parents and children associated with each disease term, the walk-hierarchy python code is used. The command to run the code is given below, where i represent the identifier (node\_code) of the disease term.

python3 walk-hierarchy.py -k YOUR\_API\_KEY -i identifier -s SNOMEDCT\_US -o parents / children

For eg., if we want to check the parents and children associated with the disease Albinism, the following commands are used.

python3 walk-hierarchy.py -k YOUR\_API\_KEY -i 15890002 -s SNOMEDCT\_US -o parents

python3 walk-hierarchy.py -k YOUR\_API\_KEY -i 15890002 -s SNOMEDCT\_US -o children

15890002 is the node code of Albinism which is taken from the condensed knowledge graph.

Node codes of all disease node terms are in the file node\_code\_of\_disease\_CUI.xlsx .