**Phenotypes and symptoms lexicon**

**Resources:**

• Human Phenotype Ontology (HPO)

• Phenotype-Wide Association Studies (PheWAS)

• Online Mendelian Inheritance in Man (OMIM) database

We used our diseases lexicon to add missing synonyms and to identify the duplicates across the resources. Please see diseases lexicon project for details. We compiled the diseases lexicon from two resources.

• Unified Medical Language System (UMLS) Metathesaurus

• Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED - CT)

**Human Phenotype Ontology (HPO):**

URL: http://human-phenotype-ontology.github.io/

Downloaded file: hp.obo

We developed a Java program to extract the phenotypes from hp.obo.

$ javac HPOMiner.java

$ java HPOMiner hp.obo HPO\_OUTPUT

We developed a Java program to extract the resources information for every phenotype.

$ javac HPOMappingResources.java

$ java HPOMappingResources hp.obo HPO\_RESOURCES\_OUTPUT

We developed a Java program to get list of phenotypes and symptoms.

$ javac PhenotypeLister.java

$ java PhenotypeLister HPO\_OUTPUT HPO\_PHENOTYPES\_AND\_SYMPTOMS

**Phenome-wide association studies (PheWAS) for ICD-9 codes:**

URL: https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/icd-9-phewas-code-map-version-12

Version: PheWAS Code version 1.2, published in 2015

Downloaded file: phecode\_icd9\_rolled.csv

To download the file, click on “Export All” button to download the entire table. The file is saved as .csv.

To convert the file to ‘Tab delimited Text (.txt)’ use some online tools. We used onlinecsvtools (https://onlinecsvtools.com/convert-csv-to-tsv).

To group ICD 9 codes based on Phe codes

$ javac PheWASCodesAndICD9CodesGrouper.java

$ java PheWASCodesAndICD9CodesGrouper phecode\_icd9\_rolled.txt PHEWAS\_OUTPUT

**Online Mendelian Inheritance in Man (OMIM) database:**

URL: https://www.omim.org/phenotypicSeriesTitle/all

Downloaded file: Phenotypic-Series-Titles-all.txt

**Phenotypes Lexicon from HPO, PheWAS and OMIM:**

$ javac PhenotypesLexiconGenerator.java

$ java PhenotypesLexiconGenerator ~/HPO\_PHENOTYPES\_AND\_SYMPTOMS ~/PHEWAS\_OUTPUT ~/Phenotypic-Series-Titles-all.txt PHENOTYPES\_LEXICON\_FROM\_HPO\_PHEWAS\_OMIM

**Resources for prefixes and suffixes for medical terms:**

Reference: https://en.wikipedia.org/wiki/List\_of\_medical\_roots,\_suffixes\_and\_prefixes

File: prefix\_suffix\_for\_medicalTerms.txt

**To remove prefixes and suffixes from phenotype names:**

We developed a Java program to remove prefix and suffix from phenotype names. Each phenotype name is matched for all 500 prefixes and suffixes. Each processed name is saved to file with customized ID suffixed with "<underscore>D". Since, these names are derived from the original phenotype name, they are not assigned with any resource ID. The customized ID helps to identify the original phenotype name and "<underscore>D" indicated that it is a derived phenotype.

Removal of prefix and suffix can get entirely different phenotypes.

$ javac PrefixSuffixRemover.java

$ java PrefixSuffixRemover prefix\_suffix\_for\_medicalTerms.txt STANDARD\_LIST\_OF\_STOPWORDS PHENOTYPES\_LEXICON\_FROM\_HPO\_PHEWAS\_OMIM PHENOTYPES\_LEXICON\_FROM\_HPO\_PHEWAS\_OMIM\_PREFIX\_SUFFIX\_REMOVAL\_DFLAG\_OUTPUT

KinderMiner 2.0 to identify phenotypes mentioned in PubMed abstracts:

All the terms derived by removing prefixes and suffixes are not phenotypes. We used KinderMiner 2.0 our recently developed text mining tool to identify phenotypes mentioned at least in one PubMed abstracts. The process removed many false positives. Let us say PHENOTYPES\_LEXICON\_KINDERMINER\_FILTERED\_OUTPUT is the output file.

See KinderMiner project <https://github.com/stewart-lab/KinderMiner_2> for details.

**Diseases lexicon for assigning unique concept identifier (CUI):**

$ javac CUIAssigner.java

$ java CUIAssigner DISEASES\_LEXICON PHENOTYPES\_LEXICON\_KINDERMINER\_FILTERED\_OUTPUT PHENOTYPES\_LEXICON\_KINDERMINER\_FILTERED\_OUTPUT\_CUI\_ASSIGNED

**Diseases lexicon for retrieving the missing synonyms:**

It is possible that phenotypes, symptoms and derived phenotypes may have more synonyms in disease lexicon compiled from UMLS Metathesaurus and SNOMED CT. We retrieved such synonyms based on CUI assigned in the previous step. Please note that the synonyms in the original list are from Human Phenotype Ontology (HPO).

$ javac UMLSSynonymsRetriever.java

$ nohup java UMLSSynonymsRetriever PHENOTYPES\_LEXICON\_KINDERMINER\_FILTERED\_OUTPUT\_CUI\_ASSIGNED PHENOTYPES\_LEXICON\_KINDERMINER\_FILTERED\_OUTPUT\_CUI\_ASSIGNED\_SYNONYMS\_ADDED

**PSEUDO CUI:**

We manually annotated pseudo CUI for the derived phenotypes without CUI. We verified that these derived terms are actually phenotypes.