Biobank Extraction Scripts

These are the programming scripts that we used to extract the data.

Getting Started

Prerequisites

The program is currently running on Python 2.7. The following are a list of packages that were used to help with extraction. Any further requirements within each package will be installed along with package itself.

Python

- Name: pandas
  - Version: 0.20.2
  - Requires: numpy, python-dateutil, pytz
- Name: csv
- Name: ast

DataFiles

- **app15860_standard_data_2016Nov19.txt**
  - Biobank Data (Main File)
  - Contains: Phenotypic Attributes, Disease Codes (ICD10, ICD9, OPCS), Self-Reported Disease Codes (and age at diagnosis), and more
- **app15860_fixed.txt**
  - Fixed Biobank Data (Main File)
  - The above file had an issue where some of the values were squished into the same column. This

file has the issue fixed.

- **country_codes.txt**
  - Translates Hospital Location Codes to Countries
- **hes_main_sec_diag_translated.csv**
  - This is a processed file made up of, the following original, raw biobank datafiles:
    - app1372_dbtable_hesin_2017aug22.tsv (HES Main Codes File)
    - app1372_dbtable_hesin_diag9_2017aug22.tsv (HES Secondary ICD10 Codes File)
    - app1372_dbtable_hesin_diag10_2017aug22.tsv (HES Secondary ICD9 Codes File)
    - app1372_dbtable_hesin_oper_2017aug22.tsv (HES Secondary OPCS Codes File)
  - Because dates only for the HES Main Codes File, we combined with the other codes files to estimate the date at which a patient was diagnosed.
  - Columns in Processed DataFile:
    - Patient ID: eid
    - Hospital Record ID: record_id
    - Date Columns: (admidate, disdate, opdate, epistart, epiend)
    - ICD10: (diag_icd10, diag_icd10_1, diag_icd10_2, …)
    - ICD9: (diag_icd9, diag_icd9_1, diag_icd9_2, …)
    - OPCS: (oper4, oper4_1, oper4_2, …)
  - Each record_id in the file is unique, that is - no two rows can have the same record_id. However, patients CAN have multiple hospital records.
  - The file is translated to match the patient IDs

in our main UKB file.
    - Using the year of birth from the main UKB file, the program creates two years based on the age at which a patient was diagnosed.
        - icd_year: Uses the first available date in HES to calculate age, except for ‘opdate’ (date of operation)
        - opcs_year: Uses the first available date in HES to calculate age
- **hes_main_sec_apr_1997.csv**
    - This file is similar to the above but we do not diagnoses prior to April 01, 1997.

Python Main Scripts

- **CHD_Classification.py**
    - This is the program we used to classify CHD. It incorporates both the main UKB file and the HES file.
- **attrib_main.py**
    - Pulls data from UK Biobank fields based on certain extraction criteria
- **EarliestAge_Disease.py**
    - Calculates the earliest age at which a patient was diagnosed with a disease based on icd codes, opcs codes, and self-reported codes. It incorporates both the main UKB and the HES file.

Python Helper Scripts

- **main_func.py**
    - Functions that can be used for the main UKB file and the HES file (if processed by row)
- **sheets.py**
    - Functions used for extracting codes from a

```
        pandas Excelfile.
```

- **op_sheets.py**
  - File used for reading in codes from ExcelFile. It is used in conjunction with CHD_Classification.py

Python Miscellaneous Scripts

- **op_sheets.py**
  - File used for reading in codes from ExcelFile. It is used in conjunction with CHD_Classification.py
- **Quality_Check.py**
  - Useful to compare the counts of two separate extracts. Should be used after every extract before analysis.

Contributors

- **Praneetha Potiny**
- **Priyanka Saha**
- **James R. Priest**