**SearchSyn.py – A Search Tool for Synthetic Bundled .json Data**

**Scope:** This tool is part of a kit for more rapid CQL development. Currently the following tools are envisioned:

1. BundleSplitter.py –splits MITRE Synthea data into a filestructure suitable for CQL import
2. **SearchSyn.py** – iteratively searches for regular expressions within age specific criteria.
3. CDStoCQL.py – converts the output of CDS Author to smoothly import into CQL-ATOM.

**Goal:** We have a need for patients with known characteristics to test the logical flow of our CQL code. The MITRE synthetic patient data repository has ample patients with conditions, medications, and observations (blood pressure) to test out our CQL logical flow. However, the bundled .json data files are numerous and unwieldy. This automated tool meets the need for finding these patient in the copious data, while still retaining a realistic structure.

**Instructions and Use:**

1. Install python3, which comes with Linux.
2. Place SearchSyn.py near a directory containing the Synthea data
3. Execute by typing SearchSyn.py on the command line
   1. Enter the relative directory of the Synthea data. For example, *Synthea* or *data/MITRE* , but not /mnt/C/High\_blood\_pressure/Mitre/data , which is an absolute path. The program will output the number of .json files in the directory
   2. Enter the minimum age of the population, usually **18**.
   3. Enter the maximum age of the population
   4. Enter the first regular expression desired. **[Hh]ypertension**
      1. Regular expressions (regexes) offer much greater flexibility in both specificity and breadth over simple text strings. Positional filters, simple logic, and pattern matching can be utilized. Information on their use can be found here: <https://docs.python.org/3/library/re.html>
      2. **[Hh]ypertension** will search for both **Hypertension** and **hypertension** for example.
      3. Enter **STOP!** to end searching
   5. The program will then search for each patient’s birthday. Patients too old or too young are excluded, prior to the regular expression searches. The numbers of patients too young and too old is tabulated and displayed.
   6. Each of the remaining patient files are searched for the regular expression. If one matches, the corresponding filename will be displayed. The total number of matches is also displayed.
   7. Other regular expressions can be added. The program will only search the previously matched files (using AND logic) narrowing the patients that match all the regexes.
   8. User enters **STOP!** and the program ends.

**Example: Hypertension, myocardial infarction, and Captopril**

root@DAP-PC:~/HBP/HBP/Scripts# SearchSyn.py

MITRE Synthetic data search program

What directory to search? Synthea

There are 1173 patient files in /mnt/c/Informatic/HBP/Scripts/Synthea

Minimum patient age in years? 18

Maximum patient age in years? 80

Searching for patients between 18 and 80 years old

To find suitable patients, enter as many regular expressions (regex) as desired

When finished enter "STOP!"

Regular expression 1 for searching: [Hh]ypertension

Regex search string: [Hh]ypertension entered

Found- [Hh]ypertension -in Agusti╠ün529\_Gonza╠ülez124\_deb78d57-823a-45a0-81f2-68bdb5f9ee3a.json

Found- [Hh]ypertension -in Al123\_Rosenbaum794\_3d88b877-badc-4c1d-b74a-eb2082d6070e.json

Found- [Hh]ypertension -in Alberto639\_Delatorre612\_9dedae70-e46d-4de7-becd-dfb8eeaf27e6.json

Found- [Hh]ypertension -in Alejandra902\_Villa94\_63056f6b-4cb3-4f9b-9a62-c3817991e6a1.json

**.**

**. (trimmed for readability)**

**.**

Found- [Hh]ypertension -in Wendy149\_White193\_b855318c-9883-48fa-8044-1eadce3d1c5e.json

Found- [Hh]ypertension -in Wilda367\_Upton904\_a9bdba30-d449-4699-92ed-553d5e7a3595.json

Found- [Hh]ypertension -in Willian804\_Moen819\_439b24b4-6f25-4093-b101-47a39bd061ca.json

Found- [Hh]ypertension -in Wm806\_Jacobson885\_6d3ddc12-4ab5-4485-873f-45763c66e27c.json

Found- [Hh]ypertension -in Xavier983\_Quitzon246\_058ba250-99c8-457a-907a-ec9a04a1cd50.json

Found- [Hh]ypertension -in Yadira684\_Paucek755\_918867a5-6206-41a9-be53-4095f35ad971.json

Found- [Hh]ypertension -in Youlanda785\_Hudson301\_1b72c333-ced2-4f0e-9898-57a613ed245e.json

Found- [Hh]ypertension -in Yulanda554\_Balistreri607\_b038b96f-6203-4bd7-bb46-46511bd55105.json

Out of 1173 patients, 216 are under the minimum age

Out of 1173 patients, 168 are over the maximum age

257 patient files match regular expression- [Hh]ypertension -within the age range

Regular expression 2 for searching: [Mm]yocardial\s[Ii]nfarction

Regex search string: [Mm]yocardial\s[Ii]nfarction entered

Found- [Mm]yocardial\s[Ii]nfarction -in Faustino767\_Schiller186\_805cac74-6094-4f3a-aea5-2029537ba82e.json

Found- [Mm]yocardial\s[Ii]nfarction -in Mari╠üa\_Luisa888\_Oquendo599\_570a6573-8e8c-4fd0-9a5c-a5c66f2f64d8.json

2 patient files match regular expression- [Mm]yocardial\s[Ii]nfarction -within the age range

Regular expression 3 for searching: [Cc]aptopril

Regex search string: [Cc]aptopril entered

Found- [Cc]aptopril -in Mari╠üa\_Luisa888\_Oquendo599\_570a6573-8e8c-4fd0-9a5c-a5c66f2f64d8.json

1 patient files match regular expression- [Cc]aptopril -within the age range

Regular expression 4 for searching: STOP!

Stopping loop

root@DAP-PC:~/HBP/HBP/Scripts#