**Preparation:**

For our track, it may be useful to have UMLS license (free). It may a little time to get/be approved- so please try to get it as soon as you can: <https://www.nlm.nih.gov/research/umls/quickstart.html>

Direct link:

<https://uts.nlm.nih.gov/license.html>

You may want register for this so have api key for use at hackathon:

<http://rxnlp.com/api-key/>

**Introduction:**

Out github is currently here:

<https://github.com/NCBI-Hackathons/cteligible>

Note the repository name may change.

Slack (got to cteligible):

<https://join.slack.com/t/nihdatasciencejuly18/shared_invite/enQtMzkxMDY1NjAwNjMwLTA4MWI2ZGQwMTI0ZTcwNWMxYjQwNmEzY2I4MjI2ODhmMzQwMWIzZGVlYWJkNWNiZjcyYTgyMWZhZDZmNDlmN2I>

Github:

<https://github.com/NCBI-Hackathons/CTEligible/invitations>

**Background:**

Please read the nlp\_resources\_background in the background directory for resources/info.

The list\_of\_trials file has the list of 342 CTEP trial NCT ID’s for which data is provided for in this project.

Eligibility\_Criteria\_example shows an example of what eligibility criteria for a trial may look like. We have structured information for only a small subset of such criteria. We have parsed unstructured text for all criteria for the 342 trials.

The nlp\_POC\_pilot\_intro presntation has slides from presentation on NLP/AI and this project.

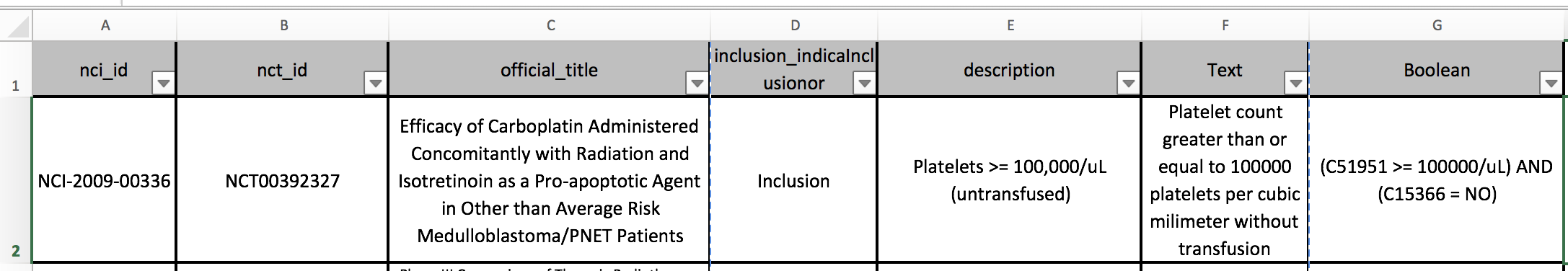
The NIH DataScience Hackathon pdf has general summary of project.

**NCI Datasets:**

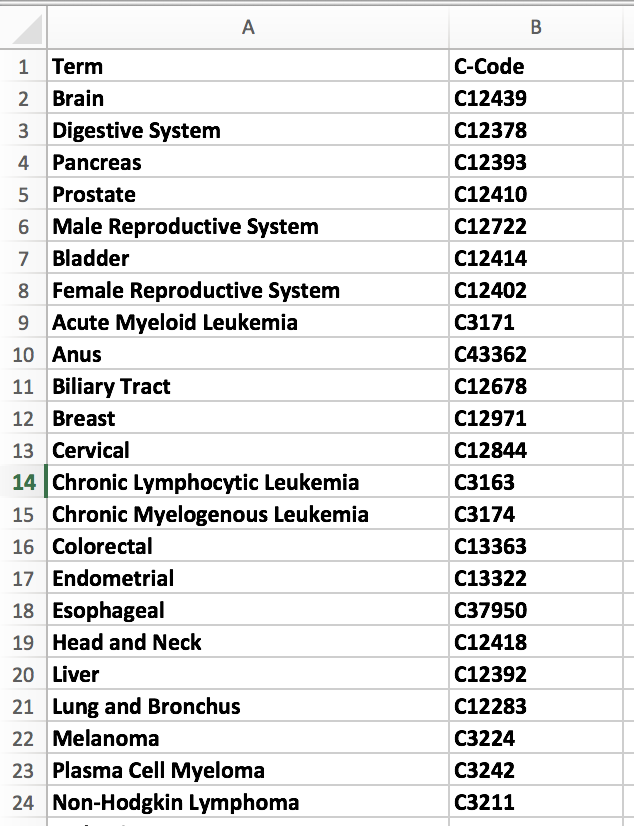
The NCI datasets are under the nci\_data directory in the repository and just released! Please see list\_of\_trials under background for the list of them and directories below for the info.

The dataset1-trials directory contains the eligibility criteria text and structured eligibility criteria from 352 CTEP trials (<https://ctep.cancer.gov/)>. When viewing in excel, it may be useful to choose “Normal” under view options. The structured eligibility criteria is encoded as logical statements with codes/display names based on NCIt (<https://ncit.nci.nih.gov/ncitbrowser/)> C-codes. Note that NCIt is also part of UMLS (and thus has associated CUI codes). The current criteria provided are: HIV, platelets, white blood cell count (WBC), and hemoglobin.

The format is as follows (nci-id, nct-id, official title, indicator if original criteria was inclusion or exclusions (note all are converted to inclusion for the final Boolean logic), description of text from original trial, text (translated description useful for encoding into boolean logic, Boolean logic (uses NCIt c-codes, mathematical relationship and Boolean operators, and UCUM for standardized units).



A second tabbed sub-sheet in the excel worksheet has the c-codes and corresponding term names as shown below.



Very important: In addition to this info, is important to note that key structured eligibility criteria for these trials is available via the CTRP API itself (and so was not uploaded). Note: It may be useful to upload to the directory so that all can access via spreadsheet. ***Please upload to the other\_data directory.*** The API is here: <https://clinicaltrialsapi.cancer.gov/>

You can see the type of information you can get (nice view in Firefox below) via this example with trial id:

<https://clinicaltrialsapi.cancer.gov/v1/clinical-trial/NCT02194738>



Relevant information that would be useful to add includes (in this priority order):

eligibility criteria->unstructured (unstructured eligibility criteria broken into one sentence at a time)

eligibility criteria-> structured (provides gender, max\_age\_in\_years, min\_age\_in\_years, etc)

sites->org\_postal\_code

phase

principal\_investigator

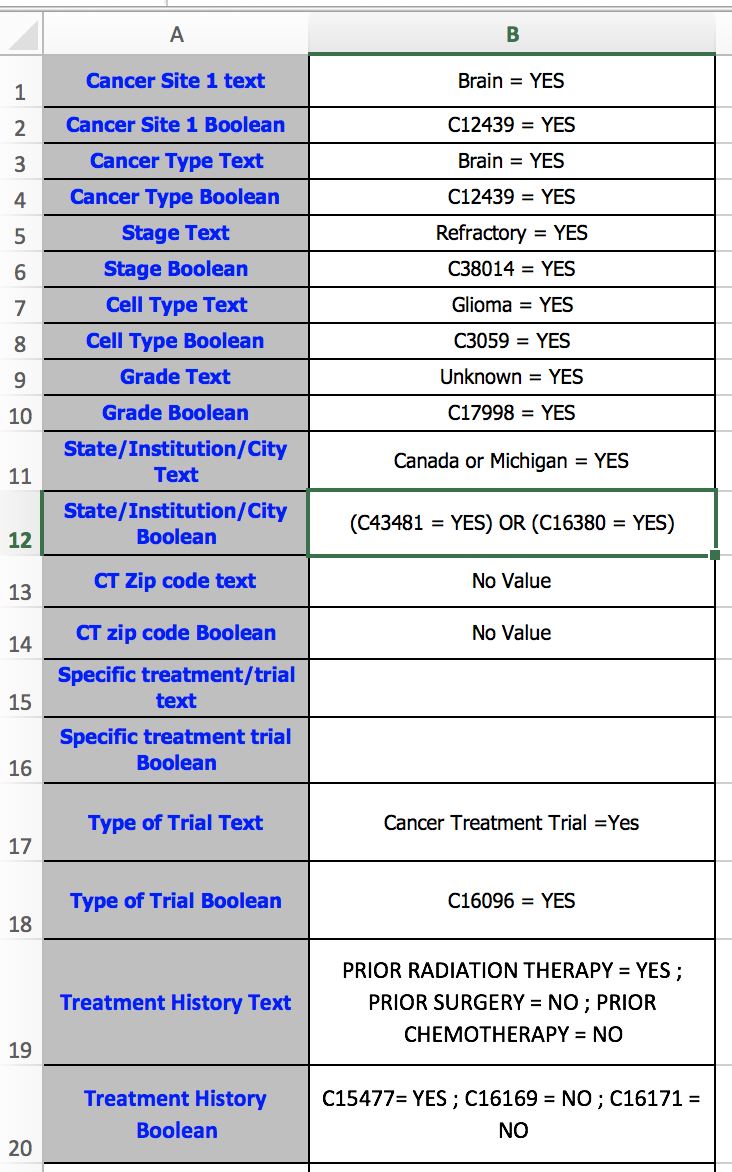
anatomic\_sites

diseases

biomarkers

API to calculate distances between zipcodes: <https://www.zipcodeapi.com/>

The dataset2-partipants directory contains 100 sampled participants who called the cancer.gov contact center. This information can be used as a source for matching to specific trials. It is pasted here in transpose (portrait mode) so that all the columns can be seen below. Basically, they alternate between text and Boolean logic representations for that text- using same approach as for dataset1-trials:



A second tabbed sub-sheet in the excel worksheet has the c-codes and corresponding term names.

**Other Datasets:**

Via PIF/NLM collaboration, we have also obtained corresponding trinary data (i.e. eligible (unconditionally), ineligible, indeterminate) for the 342 trials. That information is in: other\_data

You can read this paper for more information on trinary-type data and NLP:

<https://academic.oup.com/jamia/article/24/4/781/3038205>

Additional data from paper and code can be found here:

<https://github.com/kz26/nih-ct-hiv>

The trials\_criteria\_phrases file contains parsed criteria for the 342 trials.

The trials\_metamap\_output file contains the parsed criteria mapped to UMLS CUI’s by Metamap for the 342 trials.