ABCD Data README

Introduction

The Dosenbach Lab's ABCD collection can be found on the NIL's zfs at `/data/Daenerys/ABCD/data`. It contains the data from the NDA Collection 3165 imaging data collection (with some additional intermediary files), the raw DICOM data, and behavioral data.

Note: Not all data exists for all subjects. If there is data missing for a subject, it was either not collected or does not exist in the most recent release. We will continue to update the data on Daenerys as new data is released by the NDA.

Permission to Daenerys is restricted to individuals on the Dosenbach Lab Data Use Certificate (DUC). If you are not on the Dosenbach Lab DUC you will not be able to access the data. Both the process of being added to the DUC and being given permission to see the data take some time. In order to be given access to `Daenerys` you will need to provide us with your NIL Unix username.

Important Note: You must log out and back into your Unix account for changes in permissions to take effect. If you've been added to the DUC and are still getting a `permission denied` error when attempting to access the ABCD data after a couple days, try it in a new session. If that doesn't fix the issue, then contact us and we'll look into it.

Imaging Data

On Daenerys we have the Collection 3165 unprocessed and processed BIDS data in addition to the raw DICOM data. The DCAN Lab's GitHub has more details on the processing pipeline. For details on the BIDS format, look at the BIDS Specification.

The general folder structure is as follows:

- collection3165
    - subject_directory
        - session_directory
            - imaging_modality_directory
                - nifti run 1
                - sidecar json run 1
                - nifti run 2
                - sidecar json run 2
                - ...
    - derivatives
        - abcd-hcp-pipeline
            - subject_directory
                - session_directory
                    - imaging_modality_directory
                        - outputs

    - files
        - outputs

BIDS Format Processed and Raw Data

The raw data for each subject is found at `collection3165/sub-NDARIV*`. All processed data is found in the `derivatives` directory. The most up to date outputs from the processing pipeline are found in the `imaging_modality_directory`'s in `derivatives`. The `files` directory contains many more intermediary and output files, if the `imaging_modality_directory` doesn't have the output file you're looking for look there.

DICOM Data

The dicom data (found in the `ABCD_dicoms` directory), are kept in a BIDS-like format. In the `imaging_modality_directory` instead of a nifti for each run, there is another directory that contains the dicoms for that run.

Behavioral Data

This tutorial is intended to get you started with the wealth of non-imaging data collected on the study participants. This includes questionnaires filled out by the parents, neurocognitive batteries (e.g. IQ testing), medical information, biospecimens, etc.

Accessing the Data

The University of California San Diego curates the non-imaging data for the ABCD study. They created an interactive tool called DEAP for both exploring and performing some basic statistics on the data. Depending how sophisticated you want to get, this tool may be adequate for your needs. They also maintain a less interactive data dictionary. You will need an NIMH Data Archive account to access these tools. You can also use the ABCDE Tool to explore and access the behavioral data.

The actual data itself (release 2.0.1) is accessible at `/data/Daenerys/ABCD/data/behavioral_data/ABCDstudyDEAP`.

Data Structure

The root folder contains many files named `abcd_[shortname].txt`, each of which is a tab- and quote-delimited text file corresponding to some data `shortname`. The first row (or line) in each file is a lost of column names, e.g. `subjectkey`. The second row is a description of each column, e.g. "The NDAR Global Unique Identifier (GUID) for research subject". The second row will be helpful to you in understanding the sometimes cryptic column names from the first row, but in general you will want to skip the second row when importing the data into another program (e.g. Excel).

Each row from the the third line on corresponds to data collected from one subject at one point in time. For example:

`"subjectkey" "likes_cats" "fav_color" "unique ID" "likes cats Y or N" "favorite color"`

`"NDAR_12345" "y" "pink"`

`"NDAR_56789" "n" "blue"`

If you look under the subfolder `results/` you fill find some release notes linking each `name` to a more descriptive name describing what the data is and what the name of each field is. Even if you do not use them for analysis, you may find it more convenient to use DEAP and the data dictionary to search for files/tables and columns/fields of interest, as opposed to search throught the `results/` folder.

De-Duplicating the Data

On a simple level it is helpful to think of each file `abcd_[shortname].txt` as table of related fields (e.g. responses from the same questionnaire) where each row uniquely corresponds to one subject and each column corresponds to one response. While this mental model is useful on a general level, the reality is far more complicated.

Pitfalls

- Some subjects have contributed some of the same data twice, i.e. longitudinally.
- Some (non-longitudinal) data was collected around the same time but not on the same day.
- Some subjects filled out the same questionnaire twice on one day (not sure why study design allowed this since it's impossible to know which response is the valid one).
- Some subjects filled out some but not all of the responses on a questionnaire.
- Some questionnaires allow for contradictory responses, e.g. a subject could report report both that he has ADHD now *and* that he had it in the past but it is in remission now.

Foreign Key

If you are familiar with the SQL concept of multiple tables with foreign keys linking them together then making sense of the mess of `abcd_[shortname].txt` files will become intuitive. In an SQL table the key is a column or field that uniquely identifies a row in the table. A foreign key uniquely identifies an entity (e.g. a research subject) across multiple tables. The row number is a good *de facto* key since it is guaranteed to be unique within one table, but beware: the entries in each `abcd_[shortname].txt` file are not in the same order. Therefore the row number is not a valid foreign key for the ABCD data.

I recommend using a combination of the `subjectkey` and `interview_date` columns/fields, which will contain data that looks like `NDAR_INV0A4P0LWM` and `03/10/2018`, respectively. The `subjectkey` uniquely identifies a subject across all tables, and the `interview_date` will help you disambiguate duplicate responses due to longitudinally acquired data. Beware that a few subjects recorded multiple responses to the same question on the same data, so even with this pair of foreign keys you should still check your data for duplicates!

The Most Important Table

In general, we are only going to be intersted in subjects who have completed the MRI portion of the study

(thousands did not). Each of these subjects should have completed the entire neurocognitive battery on the same day as the MRI scan. They are enumerated in the file/table `abcd_mri01.txt` Since no subject has yet to be scanned twice, the combination of `subjectkey` and `interview_date` corresponding to the MRI scan is the unique, foreign key you most want to look for when extracting rows from other tables.

`abcd_mri01.txt` contains the following useful columns/fields:

- `subjectkey` : NDAR ID
- `interview_age` : in months at time of interview
- `interview_date` : date of interview, i.e. when non-imaging data was collected
- `mri_info_studydate` : date of MRI scan, should be identical to `interview_date`