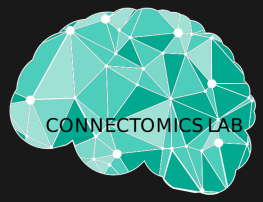
A BIDS MRI DATASET USECASE

NCCR-SYNAPSY Early-Psychosis Dataset

by Sebastien Tourbier

31 08 2021



EARLY-PSYCHOSIS DATASET

- Split into 2 datasets:
  - BIDS compliant dataset created from DICOMs: 346 subjects
  - Dataset of participants with missing DICOMs: 12 subjects

- BIDS dataset created from DICOMs validated with bids-validator:

```
$ bids-validator "/media/localadmin/HagmannHDD/Seb/PsychoDev/NiftiBIDSCompliant" --ignoreSubjectC
bids-validator@1.7.2

[...]

        Summary:                  Available Tasks:        Available Modalities:
        9212 Files, 47.25GB       rest                    MTR
        346 - Subjects                                    T1w
        6 - Sessions                                      T2map
                                                          dwi
                                                          bold
                                                          T1map
                                                          fieldmap
```

DATASET HISTORY

- Creation of pre-configured empty Datalad dataset
- Copy the content of the initial BIDS dataset with a few problems such as the use of the acquisition date as a session label.
- Edit the CHANGES

```
0.1 2021-06-05
    - Initial commit of BIDS dataset after reconversion of DICOMs to Niftis using dcm2niix
```

- Save the dataset with version tag 0.1

DATASET HISTORY

- Add and run a script in code/ folder to fix BIDS validation errors
- Validate dataset with the bids-validator
- Edit the CHANGES

```
0.2 2021-07-07
    - Create bidsify_dataset.py in code/ folder
    - Run code/bidsify_dataset.py that:
      * Add CHUV prefix to BIDS participant_id labels
      * Change session labels from acquisition date to scan1, scan2, ...
      * Add sub-<label>_sessions.tsv/.json files that stores the participant age in month and the acquisitio
      * Generate participants.tsv/.json files that stores the participant sex and the group
```

- Save the dataset with version tag 0.2
- Publish the dataset to our instutional data storage accessible via **ssh**

DATASET AVAILABILITY

Until now

- Dataset can be installed/cloned with:

```
$ datalad clone ssh://<user>@stockage-horus.chuv.ch/:/archive/PRTNR/CHUV/RADMED/phagmann/biopsycho/NiftiBI
```

Note: one needs to be inside CHUV intranet or to use CHUV VPN and have access to stockage-horus to be able to clone the dataset

In the near future

- Dataset could be installed/cloned with:

```
$ datalad clone git@github.com:NCCR-SYNAPSY/ds-biopsycho
```

Note: everybody would be able to clone the dataset from GitHub but one still needs to be inside CHUV intranet or to use CHUV VPN and have access to stockage-horus to be able to retrieve the content of the files

TIPS, TRICKS AND DEMOGRAPHIC REQUIREMENTS

HOW TO DEAL WITH LONGITUDINAL AND MULTI-SITE STUDIES

BIDS-compliant solution

- Add an extra layer of directories and file names in the form of ses-<label> to encode multiple sessions (visits)
- Use a subject label prefix identifying the site (CHUV here)

```
├── sub-CHUVA001
│   ├── ses-scan1
│   │   ├── anat
│   │   │   ├── sub-CHUVA001_ses-scan1_run-1_T1w.json
│   │   │   ├── sub-CHUVA001_ses-scan1_run-1_T1w.nii.gz
│   │   │   ├── ...
```

Reference: Longitudinal and multi-site studies section of the specifications

HOW TO DEAL WITH SIMILAR MRI ACQUIRED DURING THE SAME SESSION

BIDS-compliant solution

- Use BIDS _run-<label>_ entity to differenciate scans with same acquisition parameters
- Use BIDS _acq-<label>_ entity to differenciate similar scans with different acquisition parameters

```
├── sub-CHUVA022
│   ├── ses-scan1
│   │   ├── anat
│   │   │   ├── sub-CHUVA022_ses-scan1_run-1_T1w.json
│   │   │   ├── sub-CHUVA022_ses-scan1_run-1_T1w.nii.gz
│   │   │   ├── sub-CHUVA022_ses-scan1_run-2_T1w.json
│   │   │   ├── sub-CHUVA022_ses-scan1_run-2_T1w.nii.gz
│   │   │   ├── ...
│   │   ├── dwi
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-dsiNdir129_run-1_dwi.bval
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-dsiNdir129_run-1_dwi.bvec
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-dsiNdir129_run-1_dwi.json
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-dsiNdir129_run-1_dwi.nii.gz
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-hardiNdir147Bval3000_run-1_dwi.bval
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-hardiNdir147Bval3000_run-1_dwi.bvec
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-hardiNdir147Bval3000_run-1_dwi.json
│   │   │   ├── sub-CHUVA022_ses-scan1_acq-hardiNdir147Bval3000_run-1_dwi.nii.gz
```

Reference: Entities section of the specifications

REQUIREMENTS FOR DEMOGRAPHIC INFORMATION

- At least participant_id, age, sex and group MUST BE REPORTED
- If the study IS longitudinal:
  - participant_id, sex and group MUST be reported in the participants.tsv file
  - age MUST be reported in the session.tsv file
- If the study IS NOT longitudinal:
  - participant_id, age, sex, age, and group MUST be reported in the participants.tsv file
- Age MUST BE REPORTED in months to account for cases with multiple acquisition sessions in a given year
- All columns (except participant_id) MUST BE DESCRIBED in the participants.json / session.json

SOME OTHER RECOMMENDATIONS

- Never keep the acquisition date in the session label: this might enable the identification of the scanned individual even if MRI was anonymized or de-identified

- Keep the README of the dataset the most informative: it is the first information a user will have access to.

- De-identify MRI with a defacing tool such as pydeface if the intent is to release the dataset publicly.

  Please check this paper (Athena E. Theyers et al, 2021) for a comparison of multiple defacing tools.

- Install bids-validator locally if the dataset is big; the online validator is very slow. See the bids-validator Github repo for more details.