HABS Public Data Release – readme for raw neuroimaging data – 2020-07-11

This is the readme info the HABS raw neuroimaging release 2.0. This dataset includes the imaging data from HABS cycle 1 and 2, over a 5 year observation window. All data will be shared in NIFTI-1 format; DICOM files and ECAT files will not be shared.

First some naming conventions: **Indexing starts at 1.** HAB_1.0 is the baseline, and HAB_4.0 is the three-year follow-up, and HAB_6.0 is the five-year follow-up.

Each HABS visit is designated by HAB_x.0 and corresponds to data from a given year.month. Note that not all data from a given study year is collected at the same time. Generally speaking, all data is collected within 6 months. Also note that there was a subset that received imaging at 18 months (HAB_2.6).

Subject IDs are P_###### where # is capital letter or number. These IDs are used solely for the public dataset, and linkages to the original subject identifiers are kept under lock and key and are only available to the Harvard Aging Brain Study data managers. Dates are also blinded, and dates will appear to be from the future. Every subject is assigned a large random integer that is added to every date for that subject. For example, if a collection date was 01-01-2017, then the integer representation of that date is 736696, if the subject's random integer is 27202, then the subject's apparent date number would become 736696+27202 = 763898 and the date would then be rendered as 2091-06-24. Since the same offset is used within subject this has the effect of making all date differences within subject 100% accurate while obscuring the actual collection date.

All MR imaging data in this release are from two matched Siemens 3T Trio-Tim scanners (denoted as Bay3 or Bay4) located at the Martinos Center in Charlestown, MA. This designation is baked into the file names for the MR modalities.

HAB_1.0, HAB_4.0, and HAB_6.0 are the imaging intensive visits. HAB_2.0, HAB_3.0, and HAB_5.0 consist of a subset of clinical and neuropsychic assessments.

HAB_1.0 and HAB_4.0 each consisted of two separate MRI visits (Project 1 and Project 2 respectively) as well as a dynamic PiB scan an FDG or FTP scan. For HAB_6.0 this was changed to a single MRI visit, a dynamic PIB scan, and an FTP scan. GSP T1/T2, DTI, and SWI imaging were dropped from HAB_6.0. This means that you will see image files with the same visit code, but different dates.

All MR data was converted to NIFTI-1 format directly from DICOM using Chris Rorden's dcm2nii application (version April 1, 2010). For multi-frame data, the files are 4D NIFTI-1 files.

If a given modality is missing for a given subject at a given time point, this could be due to a number of factors including, data collection errors, scan time ran out, subject had to be removed from the scanner, scan quality was unacceptable, or participant missed a given visit. If data is missing, there is a reason that it is missing.

Included with the image downloads you will find both the sequence parameter pages from the MR console, and an example dicom header for each sequence.

MODALITY SPECIFICS:

T1 and T2 space imaging: There are two sets of MPRAGE images:

First is the 4xGRAPPA MEMPRAGE with matched T2_space that was predominantly collected in HABS Project 1. These are the same sequences as collected in the Genome Superstruct Project (GSP). This set of structural images has the benefit of a matched T2 space and short acquisition time but has lower image SNR. The T1 data set is approximately 2 gigabytes in size, and the T2 dataset is approximately 2 gigabytes in size.

513 observations across 289 participants

HAB_1.0	64
HAB_1.0, HAB_4.0	224
HAB_4.0	1

Second is the ADNI style MPRAGE that was collected predominantly in Project 2. This is a higher quality MPRAGE and is what HABS typically uses and publishes for morphological measures and FS defined ROI. Of note, several subjects were rolled into HABS from other projects where they had received an ADNI unaccelerated MPRAGE (ADNI 2 sequence), while new subjects were generally scanned with an ADNI 2x accelerated (GRAPPA) MPRAGE (note this is the same sequence that is used in ADNI 2 GO and the Dominantly Inherited Alzheimer's disease Network (DIAN)). We consider these sequences interchangeable and generally pool data across unaccelerated and accelerated scans. The unaccelerated scans are denoted as ADNI_1X and the accelerated scans are denoted with ADNI_2X. This dataset is approximately 4 gigabytes in size.

696 observations across 284 participants

HAB_1.0	56
HAB_1.0, HAB_2.6	3
HAB_1.0, HAB_2.6, HAB_4.0	4
HAB_1.0, HAB_2.6, HAB_4.0, HAB_6.0	27
HAB_1.0, HAB_4.0	57
HAB_1.0, HAB_4.0, HAB_6.0	125
HAB_1.0, HAB_6.0	8
HAB_4.0, HAB_6.0	3

SWI: An SWI sequence was acquired for microbleed counts and clinical reads. This dataset is approximately 2 gigabytes in size.

508 observations across 289 participants

HAB_1.0	68
HAB_1.0, HAB_4.0	219

```
HAB_4.0                    2
```

FLAIR: a 3D FLAIR sequence was acquired for analysis of white matter hyperintensities and clinical reads. This dataset is approximately 3 gigabytes in size.

706 observations across 289 participants

```
HAB_1.0                                    58
HAB_1.0, HAB_2.6                            3
HAB_1.0, HAB_2.6, HAB_4.0                   3
HAB_1.0, HAB_2.6, HAB_4.0, HAB_6.0         29
HAB_1.0, HAB_4.0                           60
HAB_1.0, HAB_4.0, HAB_6.0                 127
HAB_1.0, HAB_6.0                            7
HAB_4.0                                     2
```

DTI (30 direction): DTI data is 30 directions with 5 b0 images. bval and bvec files are included in addition to .nii files. This dataset is approximately 17 gigabytes in size.

499 observations across 287 participants

```
HAB_1.0                 65
HAB_1.0, HAB_4.0       212
HAB_4.0                 10
```

Resting State fMRI: The standard was to collect two back to back 6 minute runs in Project 1. This sequence is identical to the GSP resting state sequence and has a 3 second TR and 3mm isotropic voxels. Note that dummy volumes are included, so the total number of volumes is 124. We drop the first 4 frames (12 seconds) to allow for T1 equilibration, resulting in 6-minute runs. If there was concern about subject motion and time allowed, up to two additional resting state runs would be collected. When possible, we also collected a single 6-minute run in Project 2, so you will find that many subjects have 3 resting state runs spread across two session dates at the same visit (e.g. HAB_1.0 or HAB_4.0). This data can either be pooled or used for short term inter-session reliability analysis. This dataset is approximately 65 gigabytes in size.

The slice acquisition is interleaved, with the slice order going from 1, 3, 5,…, 47, 2, 4, 6, …, 46. This sequence starts at the bottom on the head goes up through the odd slices, and then returns to the bottom and moves to the top for the even slices.

1760 observations across 289 participants

```
HAB_1.0                                    59
HAB_1.0, HAB_2.6                            3
HAB_1.0, HAB_2.6, HAB_4.0                   5
HAB_1.0, HAB_2.6, HAB_4.0, HAB_6.0         27
HAB_1.0, HAB_4.0                          108
HAB_1.0, HAB_4.0, HAB_6.0                  84
```

HAB_1.0, HAB_6.0	2
HAB_4.0	1

PET data generalities:
All PET data was collected on a Siemens ECAT HR1 scanner using a 3D mode; 63 image planes; 15.2-cm axial field of view; 5.6-mm transaxial resolution; and 2.4-mm slice interval

Fludeoxyglucose (FDG - metabolism): FDG data was collected from 45 to 75 minutes post injection, and a single merged image across that time frame is provided. Most subjects received an FDG scan at HAB_1.0, and a subset received additional FDG scans at later visits. This dataset is approximately 1 gigabyte in size.

494 observations across 285 participants

HAB_1.0	160
HAB_1.0, HAB_2.6	8
HAB_1.0, HAB_2.6, HAB_4.0	4
HAB_1.0, HAB_2.6, HAB_4.0, HAB_6.0	17
HAB_1.0, HAB_2.6, HAB_6.0	7
HAB_1.0, HAB_4.0	37
HAB_1.0, HAB_4.0, HAB_6.0	30
HAB_1.0, HAB_6.0	22

Pittsburgh Compound B (PiB – amyloid beta): PiB data was collected as a full dynamic sequence from 0-60 minutes post injection and is reconstructed in 39 frames (8x15 seconds, 4 by 60 seconds, 27 x 120 seconds). Timing information (onset and duration in seconds) for each frame is provided. We have also created a 40-60 minute dataset for those who are only interested in generating SUVR and not DVR measures. The files for SUVR are the last 10 frames of the full dynamic sequence covering the 40-60 minute post injection time frame. Note that many SUVR PiB protocols are from 50-70 minutes post injection, so expect a small but consistent offset if you are comparing HABS SUVR PiB data to other 50-70 minute PiB SUVR data. You can use the GAAIN Centiloid dataset as a reference (full dynamic data from 0 to 70 minutes post injection) to compute the same measure for 40-60 and 50-70 in order to generate a linear transform for merging data. The 40-60 and 50-70 should be highly collinear with small offsets. Using one of our own pipelines for the FLR composite we computed the linear mapping as (1.0366•X)-0.0265 to map 40-60 measures to the 50-70 scale. The 40-60 dataset is approximately 13 gigabytes in size, and the full dynamic dataset is approximately 50 gigabytes in size.
There is a small set (18 scans) of the dynamic images that are missing either the first or second frame post injection due to inadequate count statistics. The timing information makes these cases transparent and should not notably affect dynamic modeling.

696 observations across 288 participants

HAB_1.0	66
HAB_1.0, HAB_2.6	3
HAB_1.0, HAB_2.6, HAB_4.0	7

HAB_1.0, HAB_2.6, HAB_4.0, HAB_6.0	41
HAB_1.0, HAB_4.0	63
HAB_1.0, HAB_4.0, HAB_6.0	97
HAB_1.0, HAB_6.0	10
HAB_4.0, HAB_6.0	1

Flortaucipir (FTP – PHF tau): FTP was not originally part of HABS cycle 1, but was added as soon as it became available. As such not all subjects received an FTP scan in the HAB_1.0 to HAB_4.0 time frame, while others received multiple FTP scans. The timing of FTP scans relative to other imaging acquisitions was initially irregular, and the FTP scan was assigned the visit code corresponding to the closest cognitive assessment. This means that unlike the other imaging modalities, you will find FTP scans with HAB_2.0, HAB_3.0, or HAB_5.0 designations. Additionally, you will find two different timing schemes for FTP data. As we were among the first sites to collect FTP data, there was an initial protocol, and then a refined protocol. The initial protocol corresponds to 80-100 minutes post injection and consists of four 5-minute frames. The latter protocol corresponds to 75-105 minutes post injection and consists of six 5-minute frames. The number of volumes in the NIFTI files (as well as the file names) will tell you which timing scheme was used. If you want to be fully consistent across all of the data you would use frames 2-5 from the 75-105 data. We have performed multiple tests comparing the 80-100 data to the 75-105 data, and we find no meaningful difference in SUVR measures between 80-100 vs 75-105. This dataset is approximately 3 gigabytes in size.

343 observations across 195 participants

HAB_1.0, HAB_4.0	1
HAB_1.0, HAB_4.0, HAB_6.0	2
HAB_2.0	7
HAB_2.0, HAB_4.0	5
HAB_2.0, HAB_4.0, HAB_6.0	6
HAB_2.0, HAB_6.0	1
HAB_3.0	6
HAB_3.0, HAB_4.0, HAB_6.0	2
HAB_3.0, HAB_5.0, HAB_6.0	5
HAB_3.0, HAB_6.0	6
HAB_4.0	27
HAB_4.0, HAB_5.0, HAB_6.0	5
HAB_4.0, HAB_6.0	87
HAB_5.0	1
HAB_5.0, HAB_6.0	8
HAB_6.0	26

Face Blinding: We have performed face and ear blinding on all sequences deemed to have sufficient coverage and resolution to support facial reconstruction. Currently, this includes T1/MPRAGE, T2 space, and FLAIR images. Face blinding was performed with our own algorithm. The process involved first running the SPM12 unified segmentation/normalization routine on

each image using the Blaiotta tissue priors (https://www.fil.ion.ucl.ac.uk/spm/toolbox/TPM/) as they have a larger field of view than the standard SPM12 tissue priors. Once done we used the ADNI MPRAGE data to create a group level representation of the tissue classes of our sample in template space. This, in combination with customized masks was used to generate an atlas that included labels for face (including the eyes, and editing out the nose), ears, areas outside the head, and areas outside the field of view. This template space mask was then reverse normalized into each subject's native image space and was used to create subject specific masks.

The face and ears were then blinded by randomizing all of the voxel values within the group level mask. This has the benefit of obscuring recognizable features of the face and ears without changing the intensity distribution of the images. Any voxels falling outside the face and ear masks, but deemed to be outside the head, were set to 0 to obscure any details that happened to fall outside the bounds of the masks (note this does change the distribution of intensity values, and can impact priors based on the distribution of image intensities). We also made an expanded brain mask for each subject, and did not edit any voxels falling within the expanded brain mask as defined on the subject level. We performed testing with FreeSurfer version 6 on 20 randomly chosen MPRAGE images to determine if this process had a notable effect on the FS recon-all process. The results showed that the face blinding procedure generally had less impact on the FS recon results than altering the affine matrix of the MPRAGE files to rigidly jitter the initial orientation of the image files. However, nothing is perfect, and there are many other tools for obtaining morphometric measures which may be affected. If you encounter problems with processing these face blinded data or notice issues where the face blinding is impacting brain tissue, please let us know. We expect face blinding algorithms to improve over time and as improvements are made, we will update the dataset with improved de-identification tools.

Finally, as the nose is removed from the images there may be greater than typical errors with co-registration, particularly with respect to the pitch angle. To help address this, we have included the affine matrix for each image that was computed during the SPM12 unified segmentation/normalization procedure on the unblinded data as an additional piece of information. This 4x4 affine matrix can be used to perform an affine mapping to standard MNI space. Using SPM12 as the base for how this information is encoded, the mapping would be done using matrix algebra by multiplying the Affine Matrix • the Voxel-to-World mapping affine matric stored in the nifti header. For example, you could perform the following procedure in MATLAB using the SPM12 package to perform an affine mapping (12 DoF linear warping) to MNI space:

```
gunzip P_8725AM_2043-06-22_HAB_6.0_ADNI_2X-MPRAGE.nii.gz;
h = spm_vol('P_8725AM_2043-06-22_HAB_6.0_ADNI_2X-MPRAGE.nii');
m = spm_read_vols(h);
Affine = load('P_8725AM_2043-06-22_HAB_6.0_ADNI_2X-MPRAGE_MNIaffine.reg');
h.fname = 'newimage.nii';
h.mat = Affine*h.mat;
spm_write_vol(h,m);
```

The resulting output image should now be in approximate alignment with any MNI template brain. (Note that this will only work if you are using a visualization program that properly uses

the affine matrix in the nifti header to map from voxel space to world space). Finally, a special thanks to Dr. Chris Schwarz at Mayo for sharing some of his experience and advice on the topic of Face Blinding neuroimaging data.

If additional information or clarification is needed, please contact habsdata@mgh.harvard.edu.

Our participants and our team have put in a huge effort to collect and curate this data, and we hope that you find it useful and that it helps to power new discoveries. Enjoy the dataset and remember to cite us!
Harvard Aging Brain Study (HABS - P01AG036694; https://habs.mgh.harvard.edu).

Sincerely,

Dr. Aaron P. Schultz and the Harvard Aging Brain team.