

Developing Human Connectome Project (dHCP) Guidelines for downloading data from the NDA

Warning! We are aware that navigating the data in the NDA interface is challenging. The best way to access and download the data is by precisely following the steps outlined below.

If you need help, email the NDA Help Desk at MDAHelp@mail.nih.gov and quote collection 3955.

Contents

Before you start you will need	2
Downloading data	
Imaging modality look-up table	
Brain Imaging Data Structure (BIDS)	
Information about Data	
Frequently asked questions	

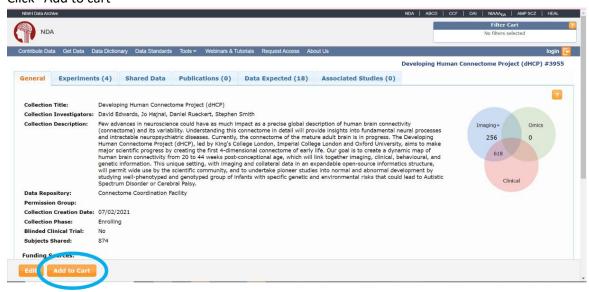


Before you start you will need

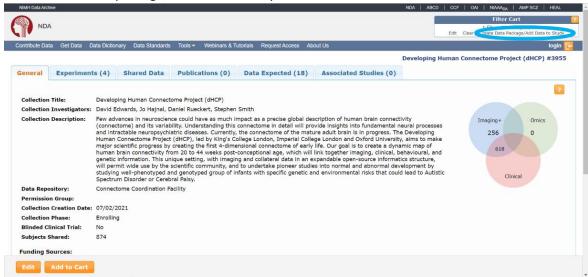
- NDA user account and login.gov account https://nda.nih.gov/nda/creating-an-nda-account.html
- 2. Data access permission https://nda.nih.gov/nda/access-data-info.html
- Download the 'Download Manager Tool' https://nda.nih.gov/nda/nda-tools.html#download-manager (see FAQ for username and password)

Downloading data

- 1. Clink on the link https://nda.nih.gov/edit_collection.html?id=3955
- 2. Click "Add to cart"

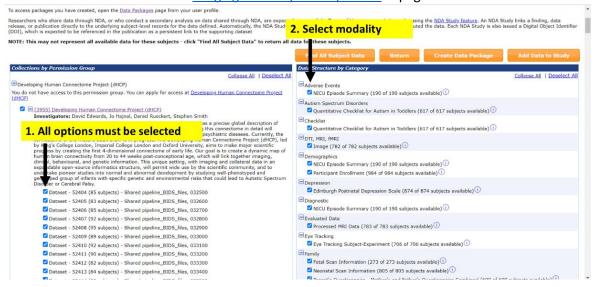


3. Click "Create Data package/Add data to study"





4. Check that all options are selected in the left table (collections by Permission Group). Select the modality in the right table ('Data Structure by Category'). The <u>list below</u> describes the data available under each heading for the 'Data structure by Category'. More information on imaging modalities can be found in the 'imaging modality look-up table' in page 8.



Adverse Events

- **NICU Episode Summary:** Clinical information for babies that stayed in the NICU. Data in file nicu101.

Autism Spectrum Disorders

- Quantitative Checklist for Autism in Toddlers: Q-CHAT scores (18 month follow-up) Scores in file qucht01.

Checklist

- Quantitative Checklist for Autism in Toddlers: Q-CHAT scores (18 month follow-up) Scores in file qucht01.

DTI, MRI, fMRI

- **Image:** Native data and motion corrected images including anatomical, dMRI, fMRI, calibrations maps and ancillary files related to the original examinations. Please also see the '<u>imaging modality look-up table'</u> in page 8.

Demographics

- **NICU Episode Summary:** Clinical information for babies that stayed in the NICU. Data in file nicu101.
- **Participant Enrollment:** Parental clinical, demographic and socio-economic information. Data in file cpenr01.

Depression

- Edinburgh Postnatal Depression Scale: EPDS scores. Data in file epds01.

Diagnostic

NICU Episode Summary: Clinical information for babies that stayed in the NICU. Data in file nicu101.

Evaluated Data

- **Processed MRI Data.** Analysed metadata for each modality (anatomical, diffusion, functional), including brain segmentations and cortical surfaces. Please also see the 'imaging modality look-up table' in page 8.

Eye Tracking

Eye Tracking Subject-Experiment: native and processed eye-tracking data. Data are in folder et_subject_experiment01. The data are organised in folders by participant. Documentation on eye-tracking task, acquisition and analysis are in folder et subject experiment01/experiments/experiment 2018/block 1/Block Design File/

Family

- **Fetal Scan Information:** Information about the fetal scan, including the gestational age on the day of the scan. Data are in file fsi01.
- **Neonatal Scan Information:** Information about the neonatal scan, including age at scan, age at birth and sex. Data are in file nnsi01.
- Parent's Questionnaire Mother's and Father's Questionnaire Combined: 18 month time point. Parenting Scale scores and parental demographics. Data are in file pqmf01.

Health

- **Neonatal Scan Information:** Information about the neonatal scan, including the gestational age at birth, the baby's age on the day of the scan and the baby's sex. Data are in file nnsi01.

Med History

- **Fetal Scan Information:** Information about the fetal scan, including the gestational age on the day of the scan. Data are in file fsi01.

Omics

 Genomics Sample: Genomic data can be found in folder genomics_sample03/Users/nickharper/cdb/dhcp_plink_files. Methylation data can be found in folder genomics_sample03/Users/nickharper/cdb/Methylation.

Parenting

- Late Pregnancy and Birth: clinical information about later stages of pregnancy and birth. Data are in file lpb01.
- Parent's Questionnaire Mother's and Father's Questionnaire Combined: 18 month timepoint. Parenting Scale scores and parental demographics. Data are in file pgmf01.

Phys Exam

- **Neonatal Scan Information:** Information about the neonatal scan, including the gestational age at birth, the baby's age on the day of the scan and the baby's sex. Data are in file nnsi01.

Questionnaire

- **Bayley-III Scales of Infant Development:** 18 month timepoint, Bayley-III scores. Data are in file bsid_iii01.
- **Child Behavior Checklist (CBCL) 1-5:** 18 month timepoint, CBCL scores. Data are in file cbcl 501.
- **Early Childhood Behavior Questionnaire:** 18 month timepoint, ECBQ scores. Data are in file ecbg01.
- Parent's Questionnaire Mother's and Father's Questionnaire Combined: 18 month timepoint. Parenting Scale scores and parental demographics. Data are in file pqmf01.
- Stimulating Parent Scale: 18 month timepoint, SPS scores. Data are in file stps01.

Resolve Identifiers

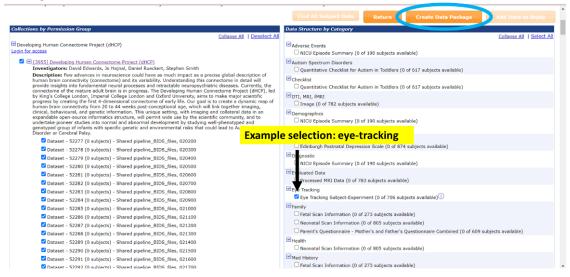
- **Genomics Subject:** List of every participant in the study for which we have genomics data. This is a subset of the Research Subject record below. Data are in file genomics_subject02.
- Research Subject: List of every participant in study. Data are in file ndar_subject01

Other

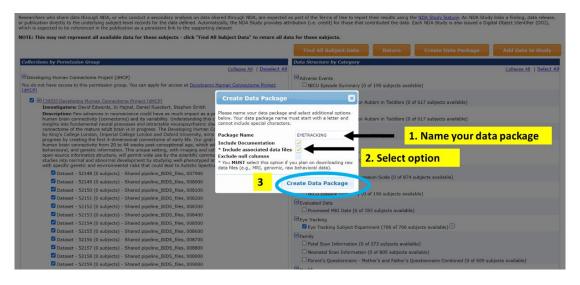
- Hammersmith Infant Neurological Examination (HINE): A 26-item examination assessing different aspects of neurological examinations such as cranial nerves, posture, movements, tone, and reflexes. Data are in file hine01.
- Fetal imaging QC: Score of image quality.



5. After selecting the modality (e.g. eye-tracking), click "Create Data Package".

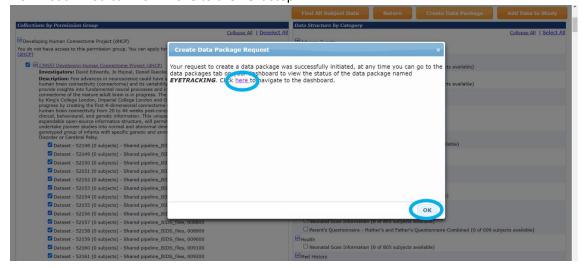


6. Create a name for your selected data package and select option "include associated files". Click "Create Data Package".

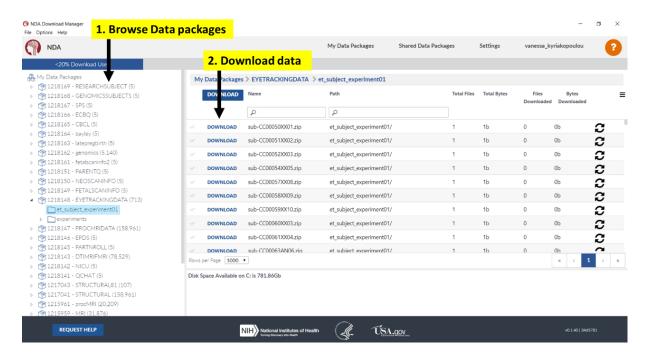




7. Popup window: Click 'here' to view the progress of the data package creation or 'ok' to stay on the package creation page. Once the data are ready to view, the status in the dashboard will change to "Ready to Download". You can now move to the next step.

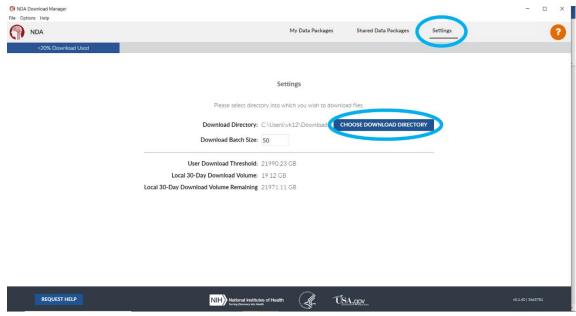


8. Open the 'Download Manager tool' and log in. The data packages are shown on the left. Navigate through the folders. The tool allows you to visualise the data available for download. Individual files can then be downloaded locally.





9. Check your download directory. Go to 'Setting' and 'choose your download directory'.



Data will be downloaded into a new folder within the download directory named "Package_nnnnnn" where "nnnnn" is the package number assigned by the NIH download process. The folder will contain the following items:

- List of NIH data structures. These will be discussed further below but all consist of the structure name with the extension ".txt", e.g. "image03.txt". They are all tab delimited text files. These structures contain scalar variables and meta data for subjects. Note that the field "interview_date" present in all structures has the North American date format "MM/DD/YYYY".
- Supporting documentation including a README.pdf and a folder containing collection documents including this document. Supporting documentation may not be present if you did not tick "Include supporting documentation" when creating the package in step 6, above.
- Package metadata: package info.txt, dataset collection txt and md5 values.txt.
- Folders containing the associated files. Present if you ticked "Include associated data files" in step 6 and if the selected structures include associated files. Structures with associated data files are image03, fmriresults01, et_subject_experiment01 and genetics_sample01. Files within the folders are arranged as the on the servers from which the data was uploaded. For example, imaging data is arranged into separate subfolders for anatomical, fmri and diffusion data.



Imaging modality look-up table

		NDA website - Data Structure by category	Folder in Download manager
Neonatal Imaging	Native data		
	Anatomical	DTI, MRI, fMRI - Image	/image03/rel3_derivatives/rel3_rawdata_vol#/sub-CC#####XX##/ses-#####/anat/
	Functional	DTI, MRI, fMRI - Image	/image03/rel3_derivatives/rel3_rawdata_vol#/sub-CC#####XX##/ses-#####/func/
	Diffusion	DTI, MRI, fMRI - Image	/image03/rel3_derivatives/rel3_rawdata_vol#/sub-CC#####XX##/ses-######/dwi/
	Calibration maps	DTI, MRI, fMRI - Image	/image03/rel3_derivatives/rel3_rawdata_vol#/sub-CC#####XX##/ses-######/B1/
			/image03/rel3_derivatives/rel3_rawdata_vol#/sub-CC#####XX##/ses-#####/fmap/
	Processed data		
	Anatomical	Evaluated Data - Processed MRI Data	/fmriresults01/rel3_derivatives/rel3_dhcp_anat_pipeline/
	Functional	Evaluated Data - Processed MRI Data	/fmriresults01/rel3_derivatives/rel3_dhcp_fmri_pipeline/
	Diffusion	Evaluated Data - Processed MRI Data	/fmriresults01/rel3_derivatives/rel3_dhcp_dmri_eddy_pipeline/
			/fmriresults01/rel3_derivatives/rel3_dhcp_dmri_shard_pipeline/
	Native data		
Fetal Imaging	Anatomical	DTI, MRI, fMRI - Image	/image03/sub-CC#####XX##/ses-######/anat/
	Functional	DTI, MRI, fMRI - Image	/image03/sub-CC#####XX##/ses-######/func/
	Diffusion	DTI, MRI, fMRI - Image	/image03/sub-CC#####XX##/ses-######/dwi/
	Calibration maps	DTI, MRI, fMRI - Image	/image03/sub-CC#####XX##/ses-######/B1/
			/image03/sub-CC#####XX##/ses-######/fmap/
	Processed data		
	Anatomical	Evaluated Data - Processed MRI Data	/fmriresults01/dhcp_anat_pipeline/
	Functional	Evaluated Data - Processed MRI Data	/fmriresults01/dhcp_fmri_pipeline/
	Diffusion	Evaluated Data - Processed MRI Data	/fmriresults01/dhcp_dmri_shard_pipeline/

Brain Imaging Data Structure (BIDS)

We have taken every effort to follow the BIDS convention (https://bids-standard.github.io/bids-starter-kit/), however not all our data are BIDS-compliant. The main reason for this is that various types of imaging data could not easily fit in existing BIDS nomenclature. Consequently, when we were designing the format for the data naming, our prime goal was to ensure that users with some experience of working with BIDS-compliant datasets find it easy to navigate our dataset. For the data dictionary, please refer to our website documentation

Information about Data

The layout of the various data structures can be found by selecting "Data Expected" on the ribbon after logging into the collection. This will show a list of all data structures together with statistics including number of subjects and number of subjects shared. There is an information button at the end of each Data Expected item. Clicking on this will generate a pop-up that lists the possible structures associated with that item. Usually there is only one possible structure that you can collect. Where options are given, select the structure with the name corresponding to the name of the structure in the download folder. After selection you will see a list of variable names, data type, description, limits on values and the meaning of categorical items. Because many of the structures are standard instruments used by NIG for multiple collections, not all the fields will be used in the dHCP project. These columns will be blank in all downloads. Ticking "Exclude null columns" when creating the package in step 6 above will allow you to ignore those fields entirely.

All data structures contain the following items:

- Subjectkey. This is the NDAR defined id that identifies the participant across all structures and across all collections.
- Src_subject_key. Alternative identifier that was used for data collection during the project. Has the form "sub-CCddddd[XX|AN|BN|CN|]ee where ddddd and ee are



- numbers. The XX, AN, BN, CN identifies singleton, twin and triplet participants. XX is used for singletons, AN for the first of twins or triplets, BN for the second of twins or triplets and CN for the third triplet. Twin and triplet participants will have the same values for ddddd and ee.
- Sex. Child's biological sex at birth: "M", "F" or "NR". There are only two participants with sex = NR. These were participants who were scanned in-utero and not seen subsequently so sex remained unknown.
- Interview_date. The date that the data in the structure was collected. These dates are randomized by subtracting a random offset from the actual date. The same offset is used for all interview_dates for each subject but different for different subjects. interview_date can be used to determine the exact time difference between two events for the same participant. It cannot be used to determine seasonality of birth.
- Interview_age and fetal_age. Interview_age is a common element in all data structures within the NIH archive. It is the age of the participant at time of data collection rounded to the nearest month. This is insufficiently precise for some analysis, e.g. image analysis; furthermore, age by this definition would be negative and possibly unknown for fetal participants scanned before birth. To address this limitation, data structures recording events that occur or could occur perinatally contain the field "fetal_age". This is the postmenstrual age in days at the time of scan. For a given participant in a data structure, only one of interview_age or fetal_age can be meaningful. Use the following logic to determine which variable to use:
 - If fetal age does not exist in the structure or if fetal_age is blank, interview_age is valid. This will be the case for the data obtained at the 18 month follow up appointment and for the genomics_sample03 structure.
 - If fetal age is present and has a non-blank value, this is the participant's post menstrual age at time of collection. Interview_age will be 0 and has no meaning.



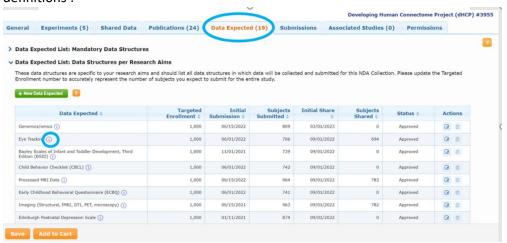
Frequently asked questions

1. Where can I find information about the study and the data available?

You can visit the study website http://www.developingconnectome.org/project/. The neonatal data publication can be found https://pubmed.ncbi.nlm.nih.gov/35677357/.

2. Where can I find a list of all variables?

All variables can be found in the NDA website, under 'Expected data structure definitions'



3. Where are the follow-up data/neurodevelopmental outcome data?

The follow-up data were collected at the 18 month time-point and can be found here:

- **Bayley-III Scales of Infant Development:** 18 month timepoint, Bayley-III scores. Data are in file bsid_iii01.
- **Child Behavior Checklist (CBCL) 1-5:** 18 month timepoint, CBCL scores. Data are in file cbcl1 501.
- **Early Childhood Behavior Questionnaire:** 18 month timepoint, ECBQ scores. Data are in file ecbq01.
- Parent's Questionnaire Mother's and Father's Questionnaire Combined: 18
 month timepoint. Parenting Scale scores and parental demographics. Data are in file
 pqmf01.
- Stimulating Parent Scale: 18 month timepoint, SPS scores. Data are in file stps01.
- Parent's Questionnaire Mother's and Father's Questionnaire Combined: 18
 month timepoint. Parenting Scale scores and parental demographics. Data are in file
 pqmf01.
- Hammersmith Infant Neurological Examination (HINE): A 26-item examination assessing different aspects of neurological examinations such as cranial nerves, posture, movements, tone, and reflexes. Data are in file hine01.
- Eye Tracking Subject-Experiment: native and processed eye-tracking data. Data are
 in folder et_subject_experiment01. The data are organised in folders by participant.
 Documentation on eye-tracking task, acquisition and analysis are in folder
 et_subject_experiment01/experiments/experiment_2018/block_1/Block_Design_Fil
 e/



Scalar variables are all found in the various data structures defined for the collection. The available structures are described in paragraph 4, above. Note that there isn't a one-to-one correspondence between the category of data shown on the download screen and the associated data structure, i.e. a given data structure may be included as part of the data for more than one category. The structure will not be duplicated in the download. When you download data as described in paragraph 5 you will find the various data structures in the downloaded folder "Package_nnnnn" where nnnnn is assigned by the download manager. Scalar data is not large so it may be helpful to download the entire collection without selecting "Include Associated Data Files" as described in paragraph 6 above. This will create a package directory with all the study scalar variables and metadata. The individual data structures are tab downloaded as tab delimited files with the extension ".txt". Note that the "interview_date" field, found in all structures, is in the North American "MM/DD/YYY" format.

5. How can I find an explanation of the meaning of the scalar variables?

Start by looking at the NIH data structure definitions on the "Data Expected" page. Further information about the specific meaning of data collected in this project can be found by visiting the study website http://www.developingconnectome.org/project/. The neonatal data publication can be found https://pubmed.ncbi.nlm.nih.gov/35677357/.

6. Why are data incomplete?

Incomplete data may be due to various reasons such as incomplete image acquisitions because the participant was unable to complete the entire scanning protocol, the baby waking up during the scan, incomplete medical records or incomplete questionnaires, participating children refusing to complete tasks etc.

7. Where are the clinical data about the perinatal period?

Clinical data about the perinatal period can here found here:

- Late Pregnancy and Birth: clinical information about later stages of pregnancy and birth. Data are in file lpb01.
- **Neonatal Scan Information:** Information about the neonatal scan, including the gestational age at birth, the baby's age on the day of the scan and the baby's sex. Data are in file nnsi01.
- **NICU Episode Summary:** Clinical information for babies that stayed in the NICU. Data in file nicu101.

8. Where can I find the age at scan for my analysis?

For neonatal data analysis, age at scan and age at birth can be found in 'Neonatal scan information' (file nnsi01). Look at variables 'NSCAN_GA_AT_BIRTH_WEEKS' and 'NSCAN_GA_AT_SCAN_WEEKS'. The variable nscan_age_at_scan_days is the participant age in days at time of scan relative to the date of birth. For fetal data analysis, gestational age at scan can be found in 'Fetal scan information (file fsi01). Look at variable 'FSCAN_GA_AT_SCAN'. These variables all report gestational (post menstrual)



age in weeks with the fractional component indicating days, e.g. 39.142857 is 39 weeks and 1 day.

9. Where can I find information about the parents?

Parental clinical, demographic and socio-economic information can be found in 'Participant enrollment' (file cpenr01). Note that the ethnicity categories are based on UK census categories. These are described in the data structure definition and do not correspond to ethnicity categories used in other countries.

10. What genomic data is there? What epigenomic data is there?

Genotyping: Salida-derived DNA samples were genotyped by NIHR BioResource Centre Maudsley Genomics & Biomarker Core Facility on the Illumina Infinium Omni5-4 array v1.2, which comprises 4327108 single-nucleotide polymorphisms (SNPs). Genotype data are provided in binary PLINK format (see https://www.coggenomics.org/plink/1.9/formats for details).

Methylation: Saliva-derived DNA from each sample was treated with sodium bisulfite (Zymo Research EZ-96 DNA Methylation Kit D5004). DNA methylation was quantified using the Illumina Infinium HumanMethylationEPIC BeadChip Kit. The raw files (.idat) are provided.

11. Where can I find the BIDS naming descriptions? Details can be found here https://bids-standard.github.io/bids-starter-kit/

12. Where can I find my username and password for the Download Manager?

Log into your NDA account and go to your profile page https://nda.nih.gov/user/dashboard/profile.html

