Data Management Plan - R2D2 - MH

Ref. Ares(2024)1164338 - 15/02/2024

1.	**Information on the Data Management Plan (DMP)**	
1.1.	Author(s) of the DMP	Sophie Dauzet, sophie.dauzet@pasteur.fr
1.2.	Date of the first version of the DMP	January 9, 2023
1.3.	Current version of the DMP and date	First version, January 9, 2023, second version 1/3/23, third version 14/02/24
1.4.	Location of storage of the DMP	Initial and intermediate and final versions will be stored on an internal server (MS Teams of the project) and published on the project website.
2.	**Information on the project**	
2.1.	Name of the project funder and funding programme	The project is funded by the European Commission under Horizon Europe funding programme
2.2.	Acronym of the project	R2D2-MH
2.3.	Title of the project	Risk and Resilience in Developmental Diversity and Mental Health
2.4.	Project ID	101057385
2.5.	Coordinator of the project	Thomas Bourgeron, Thomas.Bourgeron@pasteur.fr
2.6.	Organization and unit of the coordinator	Institut Pasteur, Human Genetics and Cognitive Functions
2.7.	Start date of the project	01/09/2020
2.8.	End date of the project	31/08/2027
3.	**Overview of the data**	
3.1.	What is the purpose of the data collection/generation?	We will acquire genetic, neurobiologica, imaging, clinical and narrative data about several to identify genetic and environmental protective/resilience factors and how they influence developmental diversity and Mental Health.
3.2.	How many dataset(s) will you generate during this project?	6
3.3.	What is the nature and format of generated/collected data?	- Genetic data - Transcriptomic data - Clinical data (questionnaire data) - Phenotypic data - Spreadsheets in .csv format as well as .docx text documents.
3.4.	Give the expected volume of generated data for this project	60 To
3.5.	Will you reuse existing data? If yes, specify their origin.	Genetic data (WGS/SNP) from cohorts such as AIMS2-Trials, SPARK, SSC, UK Biobank, dHCP and ePRIME Epigenetic data (methylation arrays) from dHCP and ePRIME Brain imaging data (MRI/EEG) from AIMS2-Trials, ABIDE, UK Biobank, dHCP and ePRIME Phenotypic/epidemiologic data from AIMS2-Trials, ABIDE, SPARK, SSC, UK Biobank, dHCP and ePRIME
4.	**Resources needed for data management**	
4.1.	What hardware resources do you need to manage your data?	Owey data lake from Institut Pasteur
4.2.	Who is in charge of data management during the research project?	responsible for data collection, processing and analysis - Thomas Bourgeron and personnel of his laboratory are responsible for the generation of the metadata and documentation related to the data - DSI Pasteur (OWEY) is responsible for data storage and responsible for data archiving and sharing

4.3.	What training or support do you think is necessary to help you manage your data?	No specific training will be necessary.
4.4.	What budget do you have for managing your data? How do you intend to cover these costs?	Institut Pasteur and the partners of the projet have a dedicated budget to provide storage and maintenance of the data internally.
5.	**Legal, ethical and security aspects**	
5.1.	Does your project include personal data?	Yes, the project includes personal data. We took legal steps to manage this type of data.
5.2.	Does your project include other data subject to a contractual, regulatory or legal obligation? If so, what type?	No
5.3.	Is there any data that should be kept confidential during your project? If so, please specify what types of data are concerned and to whom it can be made accessible	Raw data are personal data that must remain confidential and only accessible to project researchers at the Institut Pasteur. Pseudonymized intermediate results will be more widely accessible to project researchers (multi-partner project).
5.4.	What security measures are implemented for data storage during the project?	Appropriate security measures will be implemented for data: data stored on secure internal servers of each partner during the project and in OWEY, in a storage space accessible only to project researchers.
5.5.	What security measures are implemented for data collection and exchange?	Data exchange is done via OWEY, the data lake of the Institut Pasteur.
6.	**Data management during the project**	
6.1.	What is the storage location of your data during the project?	OWEY data lake of the Institut Pasteur for centralised information and: + Servers of the University of Geneva (dataset 1) + Servers of GU (dataset 3) + Servers of University of Twente and Areda repository (dataset 4) + Servers of TCD (dataset 5)
6.2.	Do you use a file classification scheme to manage your data files? Briefly indicate how it is organized.	Yes, a file classification scheme has been created for the common storage space of all project partners in OWEY. It is organized by data collection method (phenotyping, sequencing...) and then chronologically. Raw data and processed data are stored in different folders.
6.3.	What naming conventions do you use for your data? What rules do you use for clear versioning?	We are using the files format and conventions internationaly used such as for gene (VCF), brain imaging (BIDS) and phenotype (HPO).
6.4.	What measures are in place to ensure the quality of the data?	In order to guarantee the quality of the data, various measures have been implemented: - Independent repetition of the experiments (minimum of three repetitions on three different days) - Standardization of data collection - Regular review of data with PI
7.	**Data selection and long term preservation**	

7.1.	Are your data subject to preservation regulations? If yes, which ones?	The preservation constraints were defined at the time of protocol design. Data including personal data will be deleted after the publication of the last article related to this project.
7.2.	Which datasets are of long-term value and should be preserved? What are the datasets to destroy?	All datasets should be preserved due to difficulties in reproducibility and time consuming in regenerating them (except for the parts of the datasets containing personal data which will be deleted). Their preservation is essential to ensure the reproducibility of the results presented in publications and to be able to compare them with data that will be generated later.
7.3.	On which platform(s) or in which repository(s) will the datasets to be preserved be archived in the long term (after the end of the project)?	Some datasets contain sensitive data, it cannot be made available on a repository external to the Institut Pasteur. It will therefore be preserved on Institut Pasteur servers.
7.4.	Specify the formats chosen for archiving.	DICOM, BAM, CSV, SPSS, TXT…
7.5.	How long will the data be preserved?	The data will be kept for an unlimited period of time as long as the space allocated within the Institut Pasteur is available (except data including personal data which will be deleted after the publication of the last article related to this project)
7.6.	What is the expected volume of archived data?	60 To
7.7.	If a long term preservation is needed, how do you intend to cover these costs?	The costs of long term preservation will be covered by the Institut Pasteur.
8.	**Dataset 1**	
8.1.	**Data description**	
8.1.1.	Title of the dataset	Geneva Autism Cohort
8.1.2.	Who is the provider or producer of the data?	This dataset is generated by partner Université de Genève (UNIGE), Associated Partner of the project
8.1.3.	What are the nature and format of the data in this dataset?	This dataset contains clinical data stored in UNIGE secured server.
8.1.4.	Describe in more detail the data in this dataset	This dataset includes video recordings, questionnaires, results of specific developmental and cognitive assessments and interviews, eye-tracking data, electro-encephalogram, MRI and genetic data analysis.
8.1.5.	Describe the method of data collection and/or generation	Data are generated from clinical interview and observation followed by specific scoring following standards in the field, eye-tracking examinations, EEG acquisition, MRI acquisitions, blood draw and whole genome sequencing.
8.1.6.	Describe your dataset with keywords	Neurodevelopmental trajectories of children with ASD, observational interactions
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	not available at this stage of the project,

8.1.8.	What is the expected volume of data in this dataset?	about 10 To
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	No, this dataset will not be freely available because it contains personal data.
8.2.2.	If this dataset cannot be freely disseminated, explain why.	This dataset contains non-anonymized and non-pseudonymized personal data. Therefore, it cannot be made freely available to the public.
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	Due to legal constraints, this dataset will not be deposited on a data repository, it will be kept on the servers of the University of Geneva.
8.2.4.	Specify how access to this dataset will be provided in case of restriction	The dataset is stored on the servers of the University Geneva and can be accessed upon specific data transfer agreement.
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Software regularly used for this type of data: imaging (SPM, Freesurfer, cartool...), eye-tracking (Tobii Studio), and for cognitive / clinical testing mostly the Office Suite
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Not applicable, this dataset is not deposited on a data repository, so it does not have an identifier
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	The meta data are entered directly in our secured database, with a specific description for each of the data collected (author, format, date of creation, …)
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	The data are registered in FileMaker Pro, with a specific structure that allows easy retrieval of the different types of data (cognitive, clinical tests, imaging acquisitions, DNA acquisitions…)
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, comments are registered directly in the database, along with the specific test for each subject.
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, the imaging data are in dicom format. The genetic data are in bam format. Specific cognitive and clinical scores can be exported in CSV format. Dicom, bam and CSV formats are open formats and therefore interoperable.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	NA
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	we use standard vocabulary
8.5.	**Increase data reuse**	

8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published and if sufficiently anonymized, published dataset may be reused by the scientific community upon specific data transfer agreement.
8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	Not applicable, data will not be freely available
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse within 1 year, after manuscript publication.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored during at least 10 years after the data publication: This might be longer depending on storage availabilities.
8.	**Dataset 2**	
8.1.	**Data description**	
8.1.1.	Title of the dataset	Bavarian Longitudinal Study
8.1.2.	Who is the provider or producer of the data?	This dataset is held by Professor D. Wolke, Univeristy of Warwick
8.1.3.	What are the nature and format of the data in this dataset?	This dataset contains epidemiological data stored in SPSS files
8.1.4.	Describe in more detail the data in this dataset	Please see details of the dataset at: https://platform.recap-preterm.eu/pub/study/best_bls Phase I (1984-1990) Four assessments (1985-1990) were carried out daily during the neonatal period, at discharge from hospital, at 5 months and 20 months (corrected for prematurity) and at 56 months of (=4;8 years) chronological age. Study aims: a) to document the prevalence of somatic and psychosocial outcomes of neonatal problems and early delivery in an unselected population of children in Germany; and b) to evaluate the impact of a regionalised neonatal service to the locally organised service provisions in Germany. For this purpose a geographically defined region in Finland was also studied and the same instruments were used (AYLS). Phase II (1990-1997) Two assessments (1991-1993) at 6 and 8 years; Study aims: more intensive investigation of a reduced sample, assessed impact of biological and social risk on cognitive, social and behavioral development and academic achievement, intensive interviewing on parental caretaking and psychopathology. Phase III (1997-1999) One assessment (1998-1999) at 13 years (questionnaires were sent out only); Study aims: development of literacy, language and mathematical skills and school achievement, focus on the highest risk group (VP/VLBW) of the original sample only. Phase IV (2009-2015) One assessment (2010-2014) at 26 years. Study aims: identification of protective and resiliency factors in the development of children exposed to biological or environmental adversity. The study focused on neuronal, neuro-cognitive and behavioral development, health utility, and quality of life from infancy to adulthood. Phase V (Telephone interviews of the Life course were completed at age 34 years.
8.1.5.	Describe the method of data collection and/or generation	Data include standardised cognitive tests, psychiatric interviews, questionnaires of beahviour and detailled information of social functionsing (e.g. freindships and bullying). A range of potential protective factors such as friendship relationships and aprenting were collected. - again see above webpage to not repeat the same here.
8.1.6.	Describe your dataset with keywords	preterm, child, adulthood, ADHD, cognitive, school achievemnt, csocail functioning, psychioatric symptoms, population study

8.1.7.	Indicate the URL or the persistent identifier to access your dataset	not available at this stage of the project, the identifier will be indicated upon publication of the dataset. The dataset is not to be published in the foreseeable future as further data collection and exploitation under way.
8.1.8.	What is the expected volume of data in this dataset?	9 assessment waves - see above for the phases
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	No, this dataset will not be freely available because it contains personal data. Pseudoanonymised data can be made available as required for specific publications
8.2.2.	If this dataset cannot be freely disseminated, explain why.	This dataset contains non-anonymized and non-pseudonymized personal data. Therefore, it cannot be made freely available to the public.
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	Not applicable, this dataset will not be made freely available. The newly collected genetic data will be made avialble to the consortium for scientific purposes.
8.2.4.	Specify how access to this dataset will be provided in case of restriction	Usually on request and specified data sharing agreements.
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Regularly used software: excel, SPSS
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Not applicable, this dataset is not deposited on a data repository, so it does not have an identifier
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Given that no disciplinary standard exists in our field and that this dataset will not be deposited in a repository, we have developed detailled data documentation. Furthermore harmonziation has been conducted with other similar preterm cohorts a catalogue can be found at https://platform.recap-preterm.eu/pub/search#lists?type=studies&query=study(limit(0,20))
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	see at https://platform.recap-preterm.eu/pub/search#lists?type=studies&query=study(limit(0,20)) the search criteria
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, we have developed detailled documentation for most assessments (word files, PDF of instruments, excel data coding)
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, can be converted in various formats from SPSS.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We did not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in psychology, specific terms are specified in the documentation associated with the dataset.
8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	The data on genetics (anonymised), subject to participant consent, can be reused. Other data that are or were funded from elsewhere are not freely available but can be made available on request.

8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	This dataset will not be made freely available, therefore no license is associated with it. Access to the dataset will be provided upon request and the conditions of use will be defined at the time of sharing.
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	see above
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum period allowed by participant consent
8.	**Dataset 3 - WP4, Task 4 (GU samples)**	
8.1.	**Data description**	
8.1.1.	Title of the dataset	Frankfurt A-FFIP cohort & Frankfurt ADOS-cohort (included in WP4, task 4)
8.1.2.	Who is the provider or producer of the data?	Goethe Universiteit
8.1.3.	What are the nature and format of the data in this dataset?	csv or txt-files
8.1.4.	Describe in more detail the data in this dataset	data from behavioral questionnaires, direct behavior observation, parental interviews, SNP data
8.1.5.	Describe the method of data collection and/or generation	data from behavioral questionnaires, direct behavior observation, parental interviews: have been collected by clinical researchers in the related projects; data will be partially re-used and partially first collected during R2D2. SNP data: will be generated at Life and Brain in Bonn, after DNA has been obtained from the children. DNA is extracted and prepared at GU.
8.1.6.	Describe your dataset with keywords	children with Autism Spectrum Disorder, longitudinal behavioral and cognitive data, first assessment age <=5.5 years old
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	Not available yet.
8.1.8.	What is the expected volume of data in this dataset?	To be determined
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	No
8.2.2.	If this dataset cannot be freely disseminated, explain why.	The data contains personalised information (genetic data and diagnoses). They can only be shared with researchers who have sign a data transfer agreement with GU and the involved clinical sites in Germany (Würzburg, Augsburg)
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	Due to legal constraints, this dataset will not be deposited on a data repository, it will be kept on the servers of GU (Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy).
8.2.4.	Specify how access to this dataset will be provided in case of restriction	The dataset is stored on the servers of GU (Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy) and is accessible upon request at c.freitag@em.uni-frankfurt.de
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Regularly used software, which can read txt-files.
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Not applicable, this dataset is not deposited on a data repository, so it does not have an identifier
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Given that no disciplinary standard exists in our field and that this dataset will not be deposited in a repository, we have defined our own metadata that will be collected in an Excel file and associated with the data.
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset is described with 3 keywords minimum
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a README file is associated with the dataset to contextualize it and summarize the analyses performed. This file is written by the experimenter at the time of data generation and will be shared along with the data files.
8.4.	**Making data interoperable**	

8.4.1.	Are the data of this dataset technically interoperable?	Yes, csv or txt format is interoperable.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	NA
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	For the SNP data, we will use the standard vocabulary related to Illumina-Chip derived SNPs.
8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published and if sufficiently anonymized, published dataset may be reused by the scientific community upon specific data transfer agreement.
8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	This dataset will not be made freely available, therefore no license is associated with it. Access to the dataset will be provided upon request and the conditions of use will be defined at the time of sharing.
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse within 2 years, after first manuscript submission and publication.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum allowed by GU (Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy) storage allowance capacity.
8.	**Dataset 4**	
8.1.	**Data description**	**PPI App study Initiation package**
8.1.1.	Title of the dataset	The effectiveness of [Name of the app to be decided] for parents of children with ASD
8.1.2.	Who is the provider or producer of the data?	Data will be collected by the University of Twente. Producers of the data are the people participating in the trial by means of self-reported quesionnaires and usage log data.
8.1.3.	What are the nature and format of the data in this dataset?	Data on effectiveness and evaluation of the app will be quantitative (questionnaire based and usage loggin) of nature and stored in CSV format.
8.1.4.	Describe in more detail the data in this dataset	This dataset includes self-reported sociodemographic data (survey items), baseline, postintervention and followup data concerning self-reported mental-health (survey items) and app-evaluation data (survey items). Additional pseudonymized app usage data are collected by log files.
8.1.5.	Describe the method of data collection and/or generation	Survey data will be collected with Qualtrics, an online questionnaire tool. Log data will be collected by the app under investigation.
8.1.6.	Describe your dataset with keywords	[app name], [type of study], effectiveness, usage , positive psychology, mindfulness, app
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	Not available at this stage of the project, the identifier will be indicated upon publication of the dataset
8.1.8.	What is the expected volume of data in this dataset?	<1 GB
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	This dataset(s) will be made freely available at the time of pre-publication of the associated article (personal data will be removed)
8.2.2.	If this dataset cannot be freely disseminated, explain why.	Not applicable
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	This dataset will be made freely available via the Areda repository.

8.2.4.	Specify how access to this dataset will be provided in case of restriction	Not applicable
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Regularly used software: Excel, SPSS or R
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Yes, this dataset is identified by a DOI issued by the Areda repository
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	The metadata standard of the Dublin Core Metadata Initiative (DCMI, n.d.) will be used.
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset is described with 3 keywords minimum
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a README file is associated with the dataset to contextualize it and summarize the analyses performed. This file is written by the experimenter at the time of data generation and will be shared along with the data files.
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, data will be saved in CSV format
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	Not applicable
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We did not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in life sciences, specific terms are specified in the documentation associated with the dataset.
8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used for commercial purposes (a CC-BY-NC license will be associated with the dataset).
8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	This dataset will be made freely available under the CC-BY-NC license (https://creativecommons.org/licenses/by-nc/4.0/)
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse within 1 year, after manuscript submission and publication.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored without limitation.
8.	**Dataset 5**	
8.1.	**Data description**	
8.1.1.	Title of the dataset	RaDiaNT Study (NRXN1 deletion cohort)
8.1.2.	Who is the provider or producer of the data?	This dataset is generated by partner 3 (TCD) and 21 (MCRI) of the project. Combined with existing data from partner 6 (GUF) and partner 1 (IP) and iPSYCH dataset.
8.1.3.	What are the nature and format of the data in this dataset?	This dataset contains demographic, cognitive, language, clinical data and genomic stored in a Delosis or REDCap database, Excel and CSV files. Audio files for languae measure will be WAV or MP3.

8.1.4.	Describe in more detail the data in this dataset	Demographic, cognitive and behavrioural data will be collected from NRXN1 deletion carriers and relatives in the form of questionnaires, interviews and assessments. Language audio recordings will be collected also. This dataset also contains biological samples (blood and/or saliva) from NRXN1 deletion carriers and relatives (caregivers and siblings), and WGS will be used to generate genomic data.
8.1.5.	Describe the method of data collection and/or generation	Data are generated using questionnaires/surveys, cognitive and language measures, clinical phenotype through observational assessments and interviews, and biosamples (blood/saliva).
8.1.6.	Describe your dataset with keywords	Human, NRXN1 deletion carriers, relatives, clinical phenotype, cognitive and language phenotype, genomics
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	Not available at this stage of the project, the identifier will be indicated upon publication of the dataset.
8.1.8.	What is the expected volume of data in this dataset?	Total of 252 NRXN1 deletion carriers and 200 relatives. Combine datasets: TCD (59 NRXN/100 relatives); MRCI (81 NRXN/100 relatives); GUF (11 NRXN); IP (11NRXN); iPSYCH (90 NRXN)
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	No this dataset will not be freely available. This dataset will be stored at TCD and IP as part of the R2D2-MH consortium. It will be shared with the consortium. Availability of data will be dependent on the consent provided by the participant.
8.2.2.	If this dataset cannot be freely disseminated, explain why.	This dataset contains personal data. Therefore, it cannot be made freely available to the public.
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	Due to legal constraints, this dataset will not be deposited on a data repository, it will be kept on the servers of TCD and the Institut Pasteur.
8.2.4.	Specify how access to this dataset will be provided in case of restriction	The dataset is stored on the servers of TCD and the Institut Pasteur. It will be shared with the consortium. Availability and accessability of data will be dependent on the consent provided by the participant.
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Regularly used software: SPSS, Excel, R
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	Not applicable, this dataset is not deposited on a data repository, so it does not have an identifier
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	Given that no disciplinary standard exists in our field and that this dataset will not be deposited in a repository, we will define our own metadata that will be collected in an Excel file and associated with the data in the IP repository: including both general (date of creation of files, file formats) and scientific (biosample type, cognitive and behavioural phentypic measures available).
8.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, this dataset will be described with 3 keywords minimum
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, a README file will be associated with the dataset to contextualize it and summarize the analyses performed. This file will be written by the experimenter at the time of data generation and will be shared along with the data files.
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, measures will be coded/scored and data will be stored in Excel or CSV format. CSV formats are open formats and therefore interoperable.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We did not use a specific ontology but we use a uniform vocabulary throughout the dataset. The vocabulary is based on commonly used terms in life sciences, specific terms are specified in the documentation associated with the dataset.
8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Once published and if sufficiently anonymized, published dataset may be reused by the scientific community upon specific data transfer agreement. Data accessability will be dependent on participant consent.

8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	This dataset will not be made freely available, therefore no license is associated with it. Access to the dataset will be provided upon request and the conditions of use will be defined at the time of sharing based on consent.
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data will be available for reuse within 1 year, after publication/project completion.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored for the maximum allowed by TCD ethics, and Institut Pasteur storage allowance capacity.
8.	**Dataset 6**	**dHCP study (waiting for information from KCL)**
8.1.	**Data description**	
8.1.1.	Title of the dataset	The Developing Human Connectome Project (dHCP)
8.1.2.	Who is the provider or producer of the data?	This dataset is generated by Kings College London (partner no.22) and co-sponsored by Guys and St Thomas' Hospital
8.1.3.	What are the nature and format of the data in this dataset?	1. Imaging Data: Anatomical [T1 weighted (T1w) and T2 weighted (T2w)], resting state functional Magnetic Resonance Imaginging (MRI) (rsfMRI) and diffusion MRI (dMRI) images 2. Demographic, Family, and Clinical Data: - Demographic Data for Parents, Mother's Past Medical History, Mother's Obstetric History, Mental Health History, Baby Medical Details at Birth 3. Neurodevelopmental and Neurocognitive Testing at 18 Months: - Behavioral information obtained through various tests and assessments performed on the participants - These behavioral measures encompass as cognitive capabilities, emotional characteristics, personality traits, and motor function 4. Genetic Information: - The dHCP dataset incorporates genetic information to study the influence of specific genetic factors on brain development and connectivity - Genotyping is performed on the participants to identify genetic variations that may contribute to conditions like Autistic Spectrum Disorder or Cerebral Palsy 5. Format and Accessibility: - The dHCP dataset is stored in an open-source informatics structure, allowing wide use by the scientific community - The dataset is made available through the he National Institute of Mental Health Data Archive (NDA), which provides a user-friendly interface for exploring and downloading the data of interest It is important to note that the dHCP dataset has different levels of access, with some data being openly accessible and others requiring researchers to apply for restricted data access Learn more: 1. [Developing Human Connectome Project (dHCP) | The Developing Human Connectome Project](https://www.developingconnectome.org/project/) 2. [Third Data Release | The Developing Human Connectome Project](https://pubmed.ncbi.nlm.nih.gov/35677357/)

8.1.4.	Describe in more detail the data in this dataset	The dHCP dataset provides a comprehensive collection of imaging data that captures the structural and functional development of the human brain from 20 weeks gestational age to full term. The dataset includes both in utero imaging of fetal brains and postnatal imaging of preterm and term-born infants, allowing for the study of typical and atypical brain development. The dHCP has collected 1228 multimodal MRI brain datasets from 1173 fetal and/or neonatal participants, together with collateral demographic, clinical, family, neurocognitive and genomic data. This dataset provides valuable insights into the structural and functional development of the human brain during the perinatal period. The MRI data in the dHCP can be categorized as follows: 1. Structural Imaging: The dataset includes T1-weighted (T1w) and T2-weighted (T2w) images. T2w images are particularly useful for anatomical segmentation and provide the substrates for functional and diffusion analysis. The dataset includes diffusion MRI (dMRI) data, which provides information about white matter development and the structural connections in the brain. 2. Functional Imaging: The dataset includes resting-state functional MRI (rsfMRI) data. The dHCP dataset also includes accompanying metadata, such as sex, age at birth, age at scan, birthweight, head circumference, and brain injury score. Additionally, there is also collateral data, such as sociodemographic and neuropsychological outcome data, as well as genomic data. The dataset being shared with Institut Pasteur includes brain images and behavioural outcomes including the Quantitative Checklist for Autism in Toddlers (QCHAT), the Early Childhood Behaviour Questionnaire (ECBQ) and the Bayley Scales of Infant and Toddler Development (BSID). Additionally, we will be sharing eye-tracking data and parenting measures (Parenting Scale [PS] and a Parent Psychiatric History Questionnaire). Genetic data (saliva samples) will be shared when participants will be assessed in childhood and will be asked to consent for data to be shared outside the UK. The current dHCP consent form states "Anonymised DNA samples may be transferred to collaborators within the U.K for genetic analysis". Learn more: 1. [First Data Release | The Developing Human Connectome Project](https://www.developingconnectome.org/data-release/data-release-user-guide/)
8.1.5.	Describe the method of data collection and/or generation	The dHCP was a prospective, observational, cross-sectional/longitudinal study. Novel imaging methods were developed to allow fetal and neonatal scanning to optimise connectivity data from MR scans. Following consent, data were collected from mothers and parents by the administration of questionnaires and by review of the medical notes to provide patient demographics, maternal illness, medications and obstetric factors. Infants were recruited at St Thomas' Hospital, London and imaged at the Evelina Newborn Imaging Centre, Centre for the Developing Brain, King's College London, United Kingdom. The Obstetric and Neonatal clinical electronic databases of Guy's and St Thomas' NHS trust was queried to provide further detailed clinical information and to allow data for comparison and validation. Saliva samples were then collected and stored to extract DNA for investigation with imaging data. During each fetal/neonatal scan visit, mothers received completed a perinatal depression scale. Children aged between 17-24 months who were recruited to the study were invited for a neurodevelopment assessment. The main caregiver completed several questionnaires in paper format. Children completed assessments of their cognitive, language and motor abilities with a developmental psychologist. In addition, gaze behaviour was characterised with gaze tracking using a Tobii TX-300 eye tracking equipment using validated tasks that assessed visual attention, cognitive control, and social behaviour. **If parent(s) consented, in addition to the saliva sample collected at the neonatal scan, a further saliva sample was collected for future ethically approved research.**

8.1.6.	Describe your dataset with keywords	Neonatal MRI Perinatal connectome Brain development Neuroimaging data Resting-state functional MRI Diffusion MRI Structural MRI Genetic and environmental risks Autistic Spectrum Disorder Cerebral Palsy
8.1.7.	Indicate the URL or the persistent identifier to access your dataset	https://nda.nih.gov/edit_collection.html?id=3955
8.1.8.	What is the expected volume of data in this dataset?	1000
8.2.	**Making data openly accessible**	
8.2.1.	Will this dataset be freely accessible?	Yes, the dHCP is an open source open-access project.
8.2.2.	If this dataset cannot be freely disseminated, explain why.	N/A
		N/A
8.2.3.	Which data repository did you choose to store and make accessible this dataset?	Data are available through the National Institute of Mental Health Data Archive (NDA). The NDA is solely responsible for the data storage, governance, quality checks and regulated access to qualified researchers, implements robust and established organizational, physical and technological security measures to protect against accidental disclosure and prevent unauthorized use and access of the data. All requests for data are reviewed by the NDA Data Access Committee to ensure they comply to applicable laws, rules, regulations and policies and are used with the intention of advancing science and medicine. The NDA uses strict terms and condition of ethical use of data imposed by contract.
8.2.4.	Specify how access to this dataset will be provided in case of restriction	To access the data you will be required to agree to a simple data sharing agreement and will then be provided with access routes of data download. To access the data through the NDA, a user will require 1. NDA user account https://nda.nih.gov/nda/creating-an-nda-account.html 2. Data access permission https://nda.nih.gov/nda/access-data-info.html 3. Download the 'Download Manager Tool' https://nda.nih.gov/nda/nda-tools.html#download-manager (see FAQ for username and password) Instructions from the dHCP website on how to download the third data release Via NDA are outlined in a pdf provided on the website (https://github.com/BioMedIA/dHCP-release-notes/blob/master/supplementary_files/NDA_guidelines.pdf)
8.2.5.	What software is necessary to read or access the data? Do you provide the documentation or the open source code of the software?	Data are made available through the NDA. A code book is also provided to download on the website (Codebook_2020-11-23). To read or access the dHCP MRI data, you will need specific software tools that are compatible with the data format. The dHCP data are typically provided in the NIfTI (Neuroimaging Informatics Technology Initiative) format, which is a common format for storing and sharing neuroimaging data. Some softwares used to read this type of data are FSL or ITK-SNAP. Instructions from the dHCP website on how to download the third data release via NDA are outlined in a pdf (https://github.com/BioMedIA/dHCP-release-notes/blob/master/supplementary_files/NDA_guidelines.pdf). Data organisation notes are also provided via a link on the dHCP website (https://biomedia.github.io/dHCP-release-notes/organisation.html)
8.3.	**Making data findable**	
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object Identifiers)? If not, describe how data and this dataset are identified.	https://nda.nih.gov/edit_collection.html?id=3955
8.3.2.	Which metadata standards do you use? If you don't use metadata standards, outline what type(s) of metadata will be created and how.	The dHCP utilized the Brain Imaging Data Structure (BIDS) metadata standards. There were no metadata standards for the clinical or follow up data however, standardised assessments and questionnaires were used.

8.3.3.	Is this dataset described by keywords in order to make it easily findable?	Yes, the dHCP dataset is described using keywords to make it easier to find. These keywords provide a concise summary of the dataset's content and research focus, enabling researchers and users to quickly identify and locate relevant data. Some of the keywords used to describe the dHCP dataset for easier discovery include: Neonatal MRI Perinatal connectome Brain development Neuroimaging data Resting-state functional MRI Diffusion MRI Structural MRI Genetic and environmental risks Autistic Spectrum Disorder Cerebral Palsy
8.3.4.	Do you provide a supplementary documentation in order to describe more precisely your data?	Yes, supplementary documentation is provided for the dHCP dataset. The supplementary documentation provides additional information about the dataset, its contents, and how to access and download the data. The supplementary documentation called “Data organisation notes” includes details such as data acquisition, overview of the data the number the types of imaging data available (structural and functional imaging), and the accompanying metadata for each subject (such as sex, age at birth, age at scan, birthweight, head circumference, and brain injury score). It also provides information on pipelines and directory structure (https://biomedia.github.io/dHCP-release-notes/organisation.html) Additionally, a codebook is available for download on the dHCP website (http://www.developingconnectome.org)
8.4.	**Making data interoperable**	
8.4.1.	Are the data of this dataset technically interoperable?	Yes, the open-access nature allows technical interoperability, however there are certain terms and conditions in place to ensure proper usage and sharing of the data. Additionally, the NIfTI file format demonstrates strong technical interoperability within the neuroimaging community as it has been designed to be compatible with other neuroimaging formats and tools, allowing for seamless integration and data exchange.
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	N/A
8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary interoperability. If not, will you provide mapping to more commonly used ontologies?	We did not use a specific ontology but an uniform vocabulary throughout the dataset. The standard vocabulary used in the dHCP dataset encompasses terms and concepts related to brain development, neuroimaging, and connectomics.
8.5.	**Increase data reuse**	
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	Yes, the data from the dHCP can be reused by third parties. The project has a strong commitment to open science and making the data available to the research community
8.5.2.	What license will be assigned to your dataset to permit the widest reuse possible?	No licenses are required to use the data due to them being open access
8.5.3.	When will the dataset be available for reuse? If applicable, specify why and for what period an embargo is needed.	Data are currently available for reuse. Genetic data will become available for reuse outside of the UK once consent is obtained from past participants.
8.5.4.	Specify how long the dataset will remain reusable	Data will be stored during the maximum allowed by Institut Pasteur storage allowance capacity.

8.	**Dataset #... (duplicate this part and complete it for each dataset)**		
8.1.	**Data description**		
8.1.1.	Title of the dataset		Bavarian Cohort
8.1.2.	Who is the provider or producer of	This question is important in the case of	This dataset is generated by partner 1 of
8.1.3.	What are the nature and format of the	To help you, a document that shows	This dataset contains epidemiologic
8.1.4.	Describe in more detail the data in this		This dataset includes brain images
8.1.5.	Describe the method of data collection and/or generation	Indicate how the data are generated or collected: machine-generated	Data are generated by confocal microscopy and analyzed by the ImageJ
8.1.6.	Describe your dataset with keywords	We recommend that you	Drosophila larva, behavioral
8.1.7.	Indicate the URL or the persistent identifier to	Some data repositories assign persistent	Ex 1: not available at this stage of the project,
8.1.8.	What is the expected		1 To
8.2.	**Making data openly**		
8.2.1.	Will this dataset be freely accessible?	Indicate if the finalized dataset will be freely available to	Ex 1: This dataset will be made freely available at the time of pre-
8.2.2.	If this dataset cannot be freely	Some research data can not be made See the flowchart "Legal issues related to	Ex 1: This dataset contains non-
8.2.3.	Which data repository did you choose to store and	Data repositories are the best solution for To help you find the repository	Ex 1: This dataset will be made freely available via

8.2.4.	Specify how access to this dataset will be provided in case of restriction	If the dataset is stored on a repository but not freely accessible, indicate how the dataset can be accessed: access upon	Ex 1: All requests for access to data deposited in the European Genome-phenome Archive (EGA) will be verified by the Data
8.2.5.	What software is necessary to read or access the data? Do you provide	Indicate which software you use to display, read or analyze the	Ex 1: Data access requires a software developed by
8.3.	**Making data findable**		
8.3.1.	Is this dataset identified by a persistent and unique identifier such as DOI (Digital Object	The type of identifier depends mainly on the repository you choose to deposit your	Ex 1: Yes, this dataset is identified by a DOI provided by the Zenodo repository Ex 2: This
8.3.2.	Which metadata standards do you use? If you don't use metadata	Indicate what metadata is associated with the data so that the data are	Ex 1: We plan to deposit this dataset on the PRIDE repository, a repository
		For more information about metadata and	
8.3.3.	Is this dataset described by keywords in		Yes, this dataset is described with
8.3.4.	Do you provide a supplementar y	Documentatio n is text that describes the data,	Yes, a README file is associated with the
8.4.	**Making data interoperable**		
8.4.1.	Are the data of this dataset technically interoperable?	Data is interoperable if it can be easily See our practical sheet for an	Yes, the microscopy photographs are in PNG
8.4.2.	If not, what methodologies will you apply to make your data interoperable?	If your data is in a proprietary format, you can: - if possible transform it	Our data are in a format only readable by a software developed by our service. However, we

8.4.3.	Specify whether you will be using standard vocabulary for your dataset, to allow semantic inter-disciplinary	An ontology defines a common vocabulary for researchers who need to share information in a field. It	Ex 1: As our project involves medical products for human use, we used the MedDRA (Medical
8.5.	**Increase data reuse**		
8.5.1.	At the end of the project, can the data of this dataset be reused by third parties? If reuse is restricted, explain why.	If a dataset that may be of interest to the public is made freely accessible, don't hesitate to contact the Department of Communicatio	Ex 1: Once published, this dataset may be reused by the scientific community, with the restriction that the data may not be used
8.5.2.	What license will be assigned to your dataset to permit the widest reuse	A public copyright license is a legal instrument that allows the Check out this online tool to help you choose your	Ex 1: The Zenodo repository chosen for the publication of this dataset
8.5.3.	When will the dataset be available for reuse? If	You can choose not to allow the reuse of your	Data will be available for reuse within 1 year, after
8.5.4.	Specify how long the dataset will		Data will be stored during the maximum