FAMILY

“Running in the FAMILY – Understanding and predicting the intergenerational transmission of mental illness”

Grant Agreement number: 101057529

Deliverable 2.1

1st version of data management, harmonisation, and open science plan

Workpackage:	WP 2
Task:	Task 2.2
Due Date:	31st March 2023 (M6)
Actual Submission Date:	06th April 2023 (M7)
Project Dates:	Project Start Date: October 01, 2022 Project Duration: 60 months
Responsible partner:	Partner 01 EMC, 07 FCRB
Responsible author:	Ryan Muetzel (EMC), Lisanne van Houtum (EMC), Gisela Sugranyes (FCRB)
Email:	r.muetzel@erasmusmc.nl, l.vanhoutum@erasmusmc.nl, GERNEST@clinic.cat
Contributors:	P. Camprodon, M. Ortuño

Project funded by the European Commission within HORIZON-HLTH-2021-STAYHLTH-01-02: Towards a molecular and neurobiological understanding of mental health and mental illness for the benefit of citizens and patients’		
Dissemination Level		
PU	Public — fully open (automatically posted online)	X
SEN	Sensitive — limited under the conditions of the Grant Agreement	

SUMMARY

This report includes the 1st version of data management, harmonisation, and open science plan.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Purpose and Scope

According to the FAMILY Description of Action (DoA) FCRB and EMC will create a Data Management Plan based on FAIR principles. In addition to general management and harmonisation of data, the plan will also include important information regarding GDPR, and protocols for collecting and collating all necessary documentation related to data from partners, including data use/transfer and confidentiality agreements. Working closely with WP9, all documentation necessary for facilitating open publication of work from the project will be integrated into the web portal's Knowledge Base.

This report includes the 1st version of data management, harmonisation, and open science plan.

1.2 References to other FAMILY Documents

- FAMILY DoA

2 APPENDICES

- 1st version of data management plan
- 1st version of harmonisation plan
- 1st version of open science plan

ACKNOWLEDGEMENT

Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee'. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, or the European Health and Digital Executive Agency (HADEA), the SERI or the UK Research and Innovation (UKRI). Neither the European Union nor the granting authorities can be held responsible for them.

FAMILY

“Running in the FAMILY – Understanding and predicting the intergenerational transmission of mental illness”

Grant Agreement number: 101057529

1st version of data management plan

Workpackage:	WP 2
Task:	Task 2.2
Due Date:	31st March 2023 (M6)
Actual Submission Date:	06th April 2023 (M7)
Project Dates:	Project Start Date: October 01, 2022 Project Duration: 60 months
Responsible partner:	Partner 01 EMC, 07 FCRB
Responsible author:	Ryan Muetzel, Gisela Sugranyes
Email:	r.muetzel@erasmusmc.nl, GERNEST@clinic.cat
Contributors:	M. Ortuño

Project funded by the European Commission within HORIZON-HLTH-2021-STAYHLTH-01-02: Towards a molecular and neurobiological understanding of mental health and mental illness for the benefit of citizens and patients’		
Dissemination Level		
PU	Public — fully open (automatically posted online)	X
SEN	Sensitive — limited under the conditions of the Grant Agreement	

Document History:

Version	Date	Changes	From	Review
V1	17/03/2023	First version		

Open Issues

No:	Date	Issue	Resolved
1	17/03/2023	Complete appendices 1b-1i	
2	17/03/2023	Creation Data access committee	
3	27/03/2023	Development of long-term data management plan	

SUMMARY

This is the first version of the data management plan of the FAMILY consortium.

TABLE OF CONTENTS

1 INTRODUCTION

The FAMILY consortium aims to deepen our understanding of the intergenerational transmission of mental illness, with focus on mood and psychotic disorders. For this purpose, FAMILY is set to explore the mechanisms of intergenerational transmission of psychiatric disorders, from the genetic, epigenetic, environmental and neuroimaging perspectives, and build novel prediction models that encompass the familial context. Hence, FAMILY will gather the largest population and high-risk offspring cohorts along with animal models and will implement state-of-the art approaches to integrate and analyze the data. Moreover, the consortium will investigate the bioethical and social issues brought by risk prediction. Ultimately, the FAMILY consortium will help unravel the target of preventive strategies to stop the intergenerational transmission of mood and psychotic disorders by the identification of susceptibility and resilience markers.

Due to its interdisciplinary and translational nature, the FAMILY consortium will employ and generate large quantities of multimodal human data, which is subjected to strict ethical regulations and must be handled in conformity with Standard Operating Procedures (SOPs).

The purpose of this data management plan is to establish a framework that adheres to the strictest data protection standards while guaranteeing data findability, accessibility, interoperability and reuse.

1.1 Definitions, Abbreviations and Acronyms

Table 1 List of Abbreviations and Acronyms

Abbreviation/ Acronym	DEFINITION
DRE	Digital research environment
SOPs	Standard operating procedures
DTA	Data Transfer Agreement
NA	Not available

2 DATA SUMMARY

2.1 Purpose of the data collection/generation and its relation to the objectives of the project

FAMILY aims to improve causal understanding and gain prediction power of mental illnesses from the family context and address key bioethical and social issues raised by the concept of intergenerational risk transmission and risk prediction.

The specific objectives of the project are:

1. To understand the intergenerational transmission of risk:
    i. Estimate the contribution of genetic and environmental routes of intergenerational transmission of risk from parent to offspring throughout the life course.
    ii. Identify causal factors underlying genetic and environmental routes of risk transmission and resilience.
2. To predict risk of mental illness in a familial context:
    i. Identify and validate genetic, epigenetic and brain imaging biomarkers of risk or resilience to mental disease in the family.
    ii. Develop and validate a multimodal risk prediction model and a normative modelling framework to predict, at the individual level, who is at risk of developing a mental disorder.
3. To create societal impact and end-user engagement:
    i. Map and evaluate social and ethical consequences of risk prediction for clinical use.
    ii. Increase awareness and foster active engagement of families and translate new discoveries to patients and mental healthcare professionals.

Within this framework, the FAMILY consortium will be collecting multimodal human and animal-model data and analyzing it in order to establish mental disorder risk and resilience mechanisms as well as risk prediction tools and their ethical aspects.

2.2 Specify the types and formats of the data generated/collected.

a. MRI imaging data: digital (DICOM, NIfTI format).
b. Blood: biological.
c. Saliva: biological.
d. Biological data drawn from biological samples: digital (gene SNPs, miRNA).
e. Interviews: audio (mp4) or paper.
f. Questionnaires: paper.
g. Surveys: digital.

Biological data will be stored at local biobanks. Storage of analogue data will be in locked cabinets at the site where they were originally obtained and will be kept separate from personal data. Analogue and audio data will further be transcribed and digitalised. Afterwards they will be destroyed. Blood and biological samples will be stored in local biobanks. Digital data will be stored locally on secured university servers at the site where they are originally obtained and in a secured digital research environment (DRE) to allow for centralized analyses.

2.3 Specify if existing data is being reused and how.

FAMILY will make use of existing data from the population studies and high risk-offspring cohorts detailed in Table 1.

Cohort name	Partner	Cohort N	Available data	Additional Information
Population studies				
Gen R	EMC	9778	Phenotype, Genetics, epigenetics and MRI	Appendix 1a
COPSYCH/COPSAC	Region H	650	Phenotype, Genetics, epigenetics and MRI	Appendix 1b
ALSPAC	University of Bristol	15589	Phenotype, Genetics, epigenetics	Appendix 1c
MoBa	NIPH	114500	Phenotype, Genetics, epigenetics	Appendix 1d
MCS	UCL	18818	Phenotype, Genetics	Appendix 1e
UK Biobank	*NA*	70000	Phenotype, Genetics	Appendix 1f
ABCD	UMC	11878	Phenotype, Genetics, MRI	Appendix 1g
HCP	*NA*	1206	Phenotype, Genetics, MRI	Appendix 1h
PNC	*NA*	9500	Phenotype, Genetics, MRI	Appendix 1i
Familial high-risk cohorts				
BRIDGE	EMC	208	Phenotype, Genetics, MRI	

KBO	EMC	140	Phenotype, Genetics	
MARIO	EMC	500	Phenotype, Genetics	
BASIS	FCRB	60	Phenotype, Genetics, MRI	
	FIBHGM	60	Phenotype, Genetics, MRI	
LG	CHUV	389	Phenotype, Genetics, MRI	
VIA	Region H	522	Phenotype, Genetics, epigenetics, MRI	

2.4 Data origin.

The existing population and high-risk offspring cohort information have various origins as described in Table 1. The data collection within the family consortium is specified in Table 2.

Cohort name	Partner	Estimated N to collect	Types of data
BASIS	FCRB	126	Phenotype, Genetics, MRI
BASIS	FIBHGM	62	Phenotype, Genetics, MRI
LG	CHUV	150	Phenotype, Genetics, MRI
BRIDGE	EMC	450	Phenotype, Genetics, MRI
KBO	EMC	*NA*	Phenotype, Genetics
MARIO	EMC	*NA*	Phenotype, Genetics

2.5 Data size.

The size of the data cannot be precisely addressed at the moment since the data belongs to specific partners and cohorts. However, the FAMILY consortium is expected to generate a vast quantity of multidimensional data.

2.6 Data utility.

This data will be used by members of the consortium in order to elucidate the mechanisms of the intergenerational transmission of mental illness, develop risk assessment tools and study the ethical implications of risk prediction.
Further, the data produced by FAMILY will be employed by other researchers and clinicians to develop and implement ethically informed preventive strategies that strengthen resilience and improve clinical outcomes in individuals at risk.

3 FAIR MANAGEMENT OF THE DATA

3.1 Making data findable, including provision for metadata

In general, data related to FAMILY can be found at the FAMILY website (available from the 6th month since the start of the project).

3.1.1 Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

All data produced by FAMILY will be discoverable by detailed and descriptive metadata. When possible it will be associated with persistent identification mechanisms, such as DOI.

3.1.2 What naming conventions do you follow?

FAMILY will use standardized and harmonized variable names linked to metadata in a data dictionary.

3.1.3 Will search keywords be provided that optimize possibilities for re-use?

Yes.

3.1.4 Do you provide clear version numbers?

All files will be marked with explicit dates (YYYY–MM–DD) and version numbers, where appropriate.

3.1.5 Do you provide clear version numbers?

The metadata that will be created is the standard for each data type.

3.2 Making data openly accessible

3.2.1 Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out.

Since the consortium will be operating with sensitive pseudo-anonymized human data, the principles of open science must be balanced with the need for data protection and privacy. Standard Operating Procedures (SOPs) for human data are specific to each cohort, hence there can be different procedures for data sharing or data accessibility. FAMILY members who want to use pseudo-anonymized raw data will need to ask for access to the data access committee through a Data access / Publication request form (Appendix 2) and sign the appropriate data transfer agreement. All data transfer and availability procedures will be done in compliance to the guidelines of the corresponding ethics committee. Once the access is granted, it will be available to those consortium members who requested it through the DRE.

Since it is no longer confidential, processed anonymous data will be available to all consortium members.

All FAMILY results, pipelines and scientific publications will be made openly available to every consortium member through the DRE and the FAMILY intranet and will be published in open-access journals and repositories (as far as possible).

For a detailed description of the data access and result dissemination permissions see Appendix 3.

3.2.2 What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

In order to access the data, FAMILY members will need access to the DRE, where all software needed for data access and analysis will be pre-installed. These software tools will depend on data and analysis types. The use of open-source software and code will be encouraged. Documentation about the software will be included when appropriate (e.g.: when the software is developed within FAMILY and no documentation is available online so far).

3.2.3 Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible. Have you explored appropriate arrangements with the identified repository?

Where relevant, FAMILY's methodology, code and documentation will be made available on GitHub and on the FAMILY website.

3.2.4 If there are restrictions on use, how will access be provided?

See section 3.2.1.

3.2.5 Is there a need for a data access committee?

As stated in sections 2.3 and 2.4 of this document, the data analyzed within the FAMILY consortium comes from several sources. Thus, a *data access committee* will be created in order to accommodate the SOPs of each sharing party.

3.2.6 Are there well described conditions for access (i.e. a machine readable license)?

This will be explained in the appendix since it depends on the cohorts and will be overseen by the data access committee.

3.2.7 How will the identity of the person accessing the data be ascertained?

Each consortium member has a personal and non-transferable username and password for both the FAMILY intranet and the DRE.

3.3 Making data interoperable

3.3.1 Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organizations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

The DRE provides a flexible, scalable cloud-based platform where researchers have access to and can work with the data, methods and tools that are available in FAMILY. The environment is secure, self-serviced, is capable of real-time collaboration, provides data and process audit trails, and is compliant with all laws and regulations (D4LS, Feb 2017). The DRE operates on the Microsoft Azure platform, and the hardware is located within the EU. Microsoft Azure respects the intellectual property (IP) of the researcher. The DRE facilitates FAMILY researchers to collaborate on research projects in a safe, yet flexible compute and storage environment. The architecture of the DRE allows researchers to use a solution within the boundaries of data management rules and regulations. Alongside the DRE, consortium partners will have access to the Dutch National Supercomputer ("Snellius") when high performance computing is required. In situations where a partner's data privacy protections prohibit these resources to be used, singularity containers will be implemented to safeguard against pipeline and platform-dependent biases from being introduced.

3.3.2 What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable? Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Harmonization of datasets will leverage existing efforts and plans already in place with large, EU-funded consortia using similar data types and structures (e.g. LifeCycles79 and Early Cause80) which jumpstarts the harmonization process. In nonstandard cases where differing measures/instruments have been used, a team of experts from the different sites will inventory and evaluate all phenotypic data collected across the consortium in order to identify the best options for harmonization.

To facilitate sharing and long-term inter-disciplinary use of FAMILY's data, the following formats will be chosen: pdf, txt, csv, sql, dat (SPSS), RData, DICOM, NIfTI. All files will be marked with explicit dates (YYYY–MM–DD) and version numbers, where appropriate and provenance information will be documented in the Knowledge Base. FAMILY will use standardized variable names linked to metadata in a data dictionary. In cases where possible, meta data will be included inside of files (e.g. attributes within RData structures). Industry-standard data structures will be utilized for brain imaging data and standardized processing pipelines will be applied to imaging and -omics data, in many cases within Singularity containers to ensure consistent and reproducible processing is applied uniformly to all data.

3.3.3 In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Yes.

3.4 Increase data re-use

3.4.1 How will the data be licensed to permit the widest re-use possible?

Data will not be licensed. However, there will be a long-term data re-use plan at the end of the FAMILY project.

3.4.2 When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

The data will be made available for re-use as soon as it is processed, analyzed and inspected to make sure it complies with the quality standards.

3.4.3 Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why. How long is it intended that the data remains re-usable?

A long-term data (re)use model and financial plan will be developed. This will be split into two generic categories: those which do not require access to original data, and those which do require access to original data. For the former, normative models will be made available via standard open access platforms (e.g., GitHub) for broad dissemination and use. For the latter, the DRE will be utilized. Resources will be identified to maintain the storage of the data within the DRE and outline a plan for coordinating with each consortium partner on data requests. Importantly, this strategy also allows for new groups to incorporate their data into the DRE, expanding the potential of this resource.

In accordance with local guidelines and regulations, participant data will be retained for at least 10 years upon completion of the FAMILY studies at the site where they are originally obtained and preferably at the DRE (if local legislation allow transfer of data to the DRE)

3.4.4 Are data quality assurance processes described?

Each partner is responsible for the quality control of the data collected within its own cohort.
Standard quality control procedures will be applied to processed data.

4 Allocation of resources

4.1 Estimate the costs of FAIR data and describe how we intend to cover these costs

Each partner has a specific fund devoted to the cost associated to FAIR data, where expenses like DRE support, publishing in open-access and data management costs are contemplated.

4.2 Clearly identify responsibilities for data management

The data management (data handling and data analysis) in FAMILY is the responsibility of the Coordinator (EMC), integrated within WP2 (FCRB), supported by WP7 (RUMC) with respect to infrastructure and by WP8 (LU) for issues related to research ethics. Procedures are based on the data management plan developed as part of WP2 (FCRB).
Furthermore, the Consortium Agreement (CA) will define the data-related processes and operating procedures within the consortium, including access to key knowledge (WP1, Concentris). In all cases, each FAMILY partner will be responsible for the databases from cohorts that they host or collect new data from (which is the case in four familial high-risk cohorts, EMC, FCRB, FIBHGM, CHUV) as well as for keeping records of the experiments undertaken, in line with good research practice and the FAIR principles. In FAMILY, all processing of data (when local regulations allow) will be implemented on a dedicated research infrastructure and implement strictest data protection standards (GDPR). Importantly, partners from non-EU countries have a security of information agreement with the EU (USA, Switzerland, UK, and Norway). The activities will be supported by WP2 and facilitated through the DRE infrastructure (WP2/7) and web portal (WP9)

4.3 Describe costs for long term preservation

A long-term data (re)use financial plan will be developed by the 48th month of the project.

5 Data security

5.1 What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)? Is the data safely stored in certified repositories for long term preservation and curation?

Biological data will be securely stored at local biobanks and analog data will be digitalized and destroyed. All digital data will be stored and managed through the DRE. The environment is secure and compliant with all laws and regulations (D4LS, Feb 2017). Security copies will be done periodically to ensure all data is properly preserved.

6 Ethical aspects

6.1 Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics

review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

FAMILY involves human participants, including potentially vulnerable groups and individuals (children in families with intergenerational risk of mental illness).

The research involving human subjects will comply with the WMA Declaration of Helsinki, the CoE Oviedo Convention, and other international documents on human subject research. The non-EU countries being partners in tasks involving human subject research will comply with the EU regulations in addition to their country's regulation. Both for medical and social sciences research involving human participants the ethics approval will be applied for and obtained from respective research ethics committees in advance in each country.

All consortium members and researchers in FAMILY are committed to the highest research ethics and integrity standards and will conform to the applicable international and EU law, Horizon Europe standards, and to national law in the countries where the project will be carried out. Consortium members will constantly seek advice from ethics experts involved in the consortium, external ethics experts (in the Scientific and Ethical Advisory Board [SEAB]), specialised ethics departments at their institutions and national ethics bodies, compliance managers, research ethics committees and DPOs. WP8 (work package 8: ethical aspects and social consequences of intergenerational transmission of risk and prediction of mental illness) will have the responsibility to coordinate handling previously identified, as well as any new ethical issues arising from the project. Given that ethical regulations are different between consortium member states and institutions, WP8 will also take responsibility for reviewing and advising on ethics issues if brought up by any one of the partners. The Steering Committee will evaluate ethical issues and decide upon actions at least bi-annually but more frequently when needed.

6.2 Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

In the process of recruitment, every potential participant will be informed about the planned research and receive an information sheet in their language. Potential research participants will have an opportunity to ask questions about the research project and get them answered. If potential participants will agree to participate in research, they will be asked to sign an informed consent form. Participation will be entirely voluntary, free, and fully informed consent will be obtained from all research participants. For children, parental consent and child’s assent will be obtained according to the legal regulations in each country.
In case of retraction of consent by a study participant, coupled with the wish to delete his/her personal data, all personal data of the participant will be deleted. This will be handled by the site hosting the cohort in which the participant participated, where data deleted within the DRE takes immediate effect for all partners with access.

7 Literature references

NA

8 Appendices

Appendix 1 Project Proposals

A brief project proposal must be drafted that includes a.) an abstract about the project, a brief analysis plan, data required for the project (cohorts and variables), the project timeline, and the individuals who will require data access for project.

Appendix 1a. Generation R Data Requests

Access to Generation R Data for FAMILY-related projects follows a multi-step process.

1.) A brief project proposal must be drafted (see Appendix 1).
2.) The project proposal is first sent to the data access committee via WP2. It will be briefly discussed at the SC meeting,
3.) One approved by the data access committee and SC, the data access request can proceed to the Generation R MT.
4.) Once approved by the GenR MT, access can be granted to the DRE workspace with GenR data by the GenR data manager (J Krikeb) once it has been confirmed that the partner has a.) filled in the consortium agreement addendum related to data access within the consortium and b.) each user who will be granted access has signed the Generation R confidentiality agreement.
5.) Access is granted for a period of XX months. If the project has not yet completed, access can be renewed with an addendum to the project proposal.

Appendix 1b. COPSYCH/COPSAC Data Request - Pending

Appendix 1c. ALSPAC Data Request

All FAMILY partners (and their known research teams) who have signed the consortium agreement as of MM-DD-YYYY (and the data access addendum) have been added to the ALSPAC data request, and can request data access via the data access committee with a brief project proposal. See Appendix 1 for info on the project proposal. After discussion by the SC, access will be granted to ALSPAC data by WP2.

For research staff not listed in the original ALSPAC data request, an addendum must be filed. We will file addendums twice per year, unless it is urgently necessary to file an extra addendum sooner. Contact WP2 staff if you have new research staff not listed in the original ALSPAC data request. The list can be found on the intranet (in development).

Appendix 1d. MoBa Data Request - Pending

Appendix 1e. MCS Data Request - Pending

Appendix 1f. UK Biobank Data Request - Pending

Appendix 1g. ABCD Data Request - Pending

Appendix 1h. HCP Data Request - Pending

Appendix 1i. PNC Data Request - Pending

Appendix 2. Data access / Publication request form

Appendix 2. Data access / Publication request form

FAMILY

Data access /Publication request form

BACKGROUND

Please describe your scientific question for FAMILY by completing this form. The FAMILY Data Access Committee and the Steering Committee will evaluate your proposal. After approval, data access will be granted if needed.

RESEARCH PLAN

1. Project title
Please add a descriptive project title here, from which the goal/topic of the Scientific Question (SQ) should be clear.
2. Aims and objectives
Please add a short description of the general aim & hypothesis to be assessed for this SQ and explain why it is important to address this SQ, and what is the benefit of doing this in the framework of FAMILY.
3. Study design & methods
Please add a short overview of the number of subjects needed, inclusion/exclusion criteria to be applied, techniques and tools to be used, overall approach – including any risks and alternative approaches for assessing this SQ.
4. Outcomes and Link to FAMILY
Please add a comment on the expected outcomes for this SQ and how this links to the work done in the different WPs of FAMILY (and any relevant milestones/deliverables).
5. Timelines
Please add a general overview of timelines, including study start, key milestones, and expected delivery of results.

6. Cohort(s) of interest from FAMILY
Please indicate which cohort(s) and data types you are interested in. If no data is needed, please indicate NA

APPLICANT INFORMATION

1. Principal investigator
Please provide contact details for the PI for this SQ – This PI will also take responsibility for the monitoring of the work done for this SQ and reporting the results to the FAMILY Steering Committee.
2. Key team members
Please provide an overview of the key team members to be involved in solving this SQ.

I confirm to have all ethical and legal permissions for the conduct of my study / experiment.

I confirm that if data access is granted, the data will be used only for the purpose specified in this form.

In case, our study makes any incidental findings, I confirm to report them back to the FAMILY Data Access Committee who will consult with the data owner about relevant actions to be undertaken.

Date, Place Signature Principal Investigator

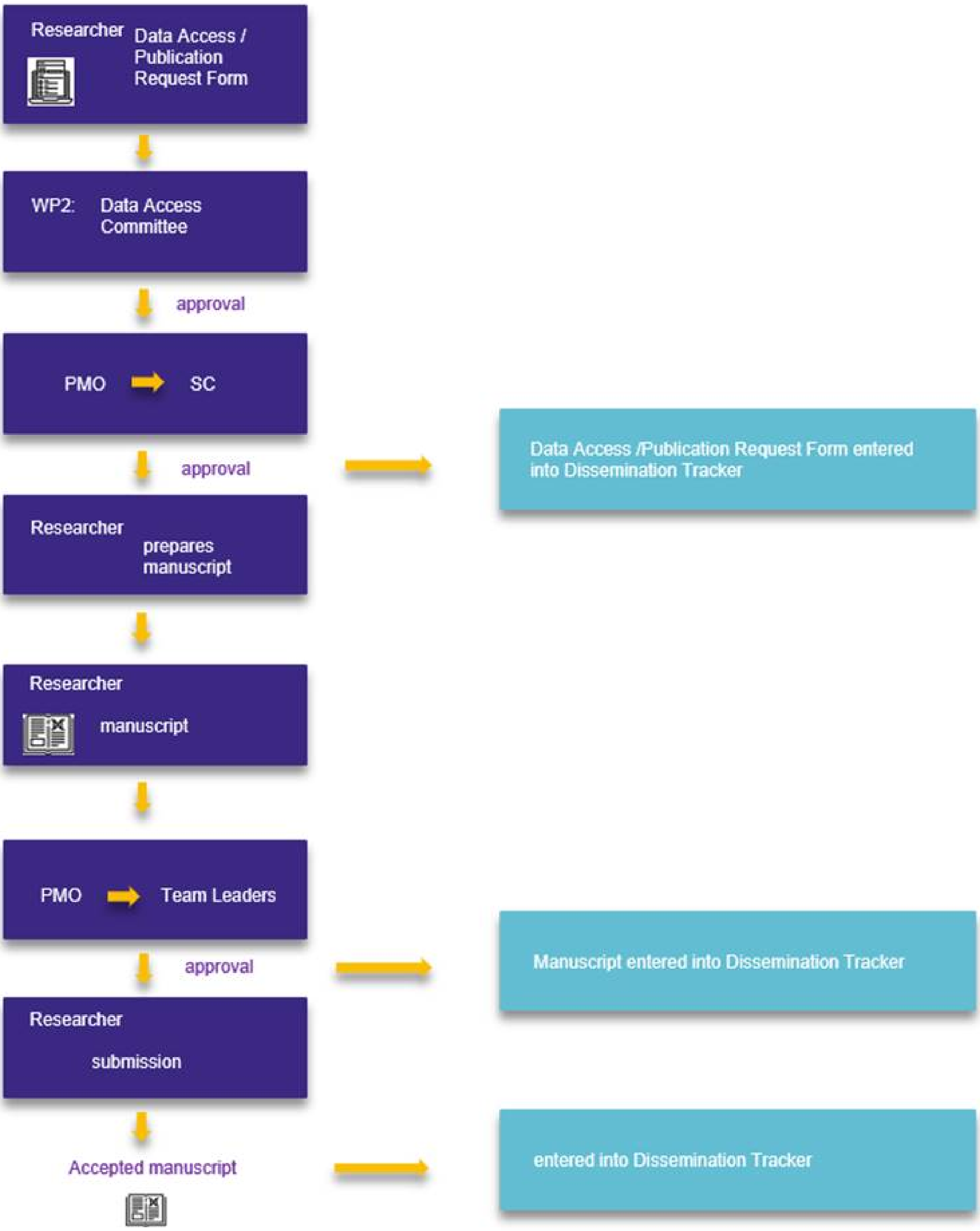
Appendix 3. Pre-submission review and approval procedure.

During the Project and for a period of 1 year after the end of the Project, the dissemination of own Results by one or several Parties including but not restricted to publications and presentations, shall be governed by the procedure of the Grant Agreement Art. 26.2 subject to the following provisions.

Prior notice of any planned publication shall be given to the other Parties at least 45 calendar days before planned submission of the manuscript to the respective journal. The main author has to follow the following procedure (Appendix figure 1):

1. The main authors, first and last authors, fill out the 'data access' or 'publication' (in case no data is necessary for the manuscript) request form (which can be downloaded from KEYWAYS) and send it to the Data Access Committee (as part of WP2) for a first eligibility check (e.g. is form filled out completely, overlap with other ongoing request, check if permissions to access/use data are in place, can research question be answered with requested data and is data available). When approved, it will be sent to the Project Management Office (PMO). The request form summarizes on 1-2 pages the proposed research, including the involved WPs, proposed authors, background, hypotheses, analyzed variables (if applicable), used methods, and key references.
2. The PMO will forward the data access/publication request form to the Steering Committee who will review it within 14 days.
3. Once approved by the Steering Committee, the planned manuscript will be entered into the FAMILY dissemination tracker by the PMO. An email will be sent to the general assembly to inform the FAMILY community of the new proposal.
4. Subsequently, authors will analyze the data (if applicable), write the manuscript, and send a final draft, that is approved by all authors, to the PMO.Created to reflect actual contributions of involved researchers.
5. The PMO informs the team leaders (representatives of an WP, an institution, or a cohort within FAMILY), who have officially 30 days to provide feedback.
6. Any objection to the planned publication shall be made in accordance with the Grant Agreement in writing to the Coordinator and to the Party or Parties proposing the dissemination within 7 calendar days after receipt of the notice. If no objection is made within the time limit stated above, the publication is permitted.

Appendix Figure 1: Pre-submission review and approval procedure


Researcher Data Access / Publication Request Form

WP2: Data Access Committee

approval

PMO → SC

approval

Data Access /Publication Request Form entered into Dissemination Tracker

Researcher prepares manuscript

Researcher manuscript

PMO → Team Leaders

approval

Manuscript entered into Dissemination Tracker

Researcher submission

Accepted manuscript

entered into Dissemination Tracker

ACKNOWLEDGEMENTS

Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee'. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, or the European Research Executive Agency (REA), the SERI or the UK Research and Innovation (UKRI). Neither the European Union nor the granting authorities can be held responsible for them.

FAMILY

“Running in the FAMILY – Understanding and predicting the intergenerational transmission of mental illness”

Grant Agreement number: 101057529

1st version of data harmonisation plan

Workpackage:	WP 2
Task:	Task 2.2
Due Date:	31st March 2023 (M6)
Actual Submission Date:	06th April 2023 (M7)
Project Dates:	Project Start Date: October 01, 2022 Project Duration: 60 months
Responsible partner:	Partner 01 EMC, 07 FCRB
Responsible author:	Ryan Muetzel, Gisela Sugranyes
Email:	r.muetzel@erasmusmc.nl, GERNEST@clinic.cat
Contributors:	P. Camprodon

Project funded by the European Commission within HORIZON-HLTH-2021-STAYHLTH-01-02: Towards a molecular and neurobiological understanding of mental health and mental illness for the benefit of citizens and patients’		
Dissemination Level		
PU	Public — fully open (automatically posted online)	X
SEN	Sensitive — limited under the conditions of the Grant Agreement	

Document History:

Version	Date	Changes	From	Review
V1	29/03/2023	First draft		

Open Issues

No:	Date	Issue	Resolved
1	29/03/2023	4.2. To compile details about particular variables within the core data domains that need to be harmonized. A data inventory will be created and completed by all partners in order to collect information of relevant variables available across the FAMILY human cohorts.	
2	29/03/2023	4.3. Defining project-specific research protocols: - 4.3.1 Genetics - 4.3.2 Epigenetics - 4.3.3 Neuroimaging: diffusion and resting state functional MRI	
3	29/03/2023	4.3.4 Harmonizing specific variables under a common format: Clinical Data	
4	29/03/2023	4.4. Assessing the quality of harmonized data	
5	29/03/2023	4.5. Preserving the data for future usage and replication	

Summary

The overarching aim of the FAMILY project is to discover and validate brain imaging, genetic, and epigenetic indicators for susceptibility or resilience to mental illness in the family. In order to accomplish this goal, the project will leverage data from both large-scale population-based cohorts, as well as high-risk clinical cohorts. Melding these different data sources together for combined analysis requires careful curation and harmonization to ensure valid results and interpretation are possible.

This document represents the first version of the Data Harmonization Plan (DHP) of the FAMILY consortium. The document outlines the steps taken to achieve harmonization, including inventorization of available data, defining minimal datasets depending on the WP objectives, establishing procedures for the harmonization of the various data types in FAMILY (e.g., clinical, imaging, genomic, epigenomic, etc.), providing harmonized sets of data to the consortium members, and documenting clearly on the knowledge base the harmonization process and, when applicable, indicating any important caveats related to the degree of harmonization.

Table of contents

1 INTRODUCTION

1.1 Purpose and Scope

This document describes the first version of the data harmonization plan. The first objective of WP 2 is to offer a secure, scalable, and existing data analysis infrastructure Digital Research Environment (DRE) with an integrative data management and harmonization plan (Task 2.1).

1.2 References to other FAMILY Documents

- FAMILY DoA

1.3 Definitions, Abbreviations and Acronyms

Table 1 List of Abbreviations and Acronyms

Abbreviation/ Acronym	DEFINITION
MRI	Magnetic Resonance Imaging
DRE	Digital Research Environment
FHR	Familial High Risk
DSM	Diagnostic and Statistical Manual of Mental Disorders
GDPR	General Data Protection Regulation
WP	Work Package
EMC	Erasmus Universitair Medisch Centrum Rotterdam
UCL	University College London
FCRB	Fundació Clínic per la Recerca Biomèdica
FIBHGM	Fundación para la Investigación Biomédica del Hospital Gregorio Marañón
CHUV	Centre Hospitalier Universitaire Vaudois
RegionH	Region Hovedstaden
NIPH	Folkehelseinstituttet
LG	Lausanne-Geneva
Gen R	Generation R
COPSYCH	Copenhagen Prospective Study on Neuro-Psychiatric Development
ALSPAC	Avon Longitudinal Study of Parents and Children
ABCD	Adolescent Brain Cognitive Development
MCS	Millennium Cohort Study
BASYS	Bipolar And Schizophrenia Young offspring Study
HCP	Human Connectome Project
KBO	Kinderen Bipolaire Ouders
MoBA	Norwegian Mother and Child Cohort Study
BRIDGE	Brain Imaging Development and Genetics
PNC	Philadelphia Neuroimaging Cohort
MARIO	Mood and Resilience in Offspring

2 Executive summary

Mental illness runs in the FAMILY. A family history of mental illness is the most important known risk factor for the development of mental health problems. Up to 50% of children with a mentally-ill parent will develop a mental disorder in their life course, suggesting a transfer of disease risk from affected parents to offspring. Such intergenerational transmission of risk of mental illness is rarely considered in clinical practice, and health care systems do not sufficiently embed family history of mental illness into diagnostics and care, leading to a delay in diagnosing patients and missing the time window for protective actions and resilience strengthening. The FAMILY project aims to discover and validate brain imaging, genetic, and epigenetic indicators for susceptibility or resilience to mental illness in the family. In order to accomplish this goal, the project will leverage data from both large-scale population-based cohorts, as well as high-risk clinical cohorts. Melding these different data sources together for combined analysis requires careful curation and harmonization to ensure valid results and interpretation are possible.

This document represents the first version of the Data Harmonization Plan (DHP) of the FAMILY consortium. One of the main objectives of FAMILY WP2 is to enhance data management process harmonization which will improve the quality and interoperability of data resources in FAMILY cohorts. The document outlines the steps taken to achieve harmonization:

- To identify and define data domains to be harmonized based on FAMILY project objectives
- To compile an inventory of relevant data across cohorts
- To define project-specific research protocols
- To harmonize specific variables under a common format
- To assess the quality of harmonized data.
- To preserve the data and associated metadata for future usage and replication.

3 General objectives

The aim of FAMILY is to understand the intergenerational transmission of risk, to predict the risk of mental illness in a familial context, and to create a societal impact and an end-user engagement. As a proof of principle, FAMILY will focus specifically on risk for mood and psychosis symptoms and diagnoses. Therefore, the FAMILY project seeks to identify and validate genetic, epigenetic, and brain imaging biomarkers for risk or resilience to mental disease in the family, and map and evaluate social and ethical consequences of risk prediction for clinical use. To achieve the established objectives, the FAMILY project will make use of a wide range of data from different modalities and samples.

Seven European Familial High-Risk (FHR) Cohorts form part of the FAMILY consortium. These samples consist of child, adolescent and young adult offspring with at least one parent with a confirmed diagnosis in the mood-psychosis spectrum (i.e. schizophrenia, bipolar disorder, depression). Family consortium partners have direct access to longitudinal information about clinical, environmental, genetic and brain imaging data.

FAMILY consortium partners have also requested access to data collected through population-based samples, which also include clinical, environmental, genetic and brain imaging data. Information about familial risk and population-based cohorts is provided in Table 1.

Population-based studies		**FHR cohorts**	
Name	**Partner**	**Name**	**Partner**
GenR/ORACLE	EMC	BRIDGE	EMC
COPSYCH	RegionH	KBO	EMC
ALSPAC		MARIO	EMC
MoBa	NIPH	BASYS	FCRB
MCS	UCL	BASYS	FIBHGM
UK Biobank		LG	CHUV
ABCD		VIA	RegionH
HCP			
PNC			

Table 1. Cohorts available to FAMILY project.

By facilitating interoperability between human cohorts and the execution of multi-cohort analyses, data harmonization and metadata curation are crucial for achieving the goals of the study. WP2 is responsible for data management in line with FAIR principles, data harmonization and GDPR-compliant storage of and access to datasets. In order to achieve this, WP2 will develop resources for the consortium that will aid in data harmonization and analysis. These resources will include a Digital Research Environment (DRE) to store and share data with other partners, a web-based Knowledge Base, which will include a detailed and searchable data dictionary of all the raw data and ethical guidelines, and a GitHub repository.

Of note, this document focuses on the consortium's harmonization of human data (it does not describe harmonization of data obtained from animal models within the Consortium). In addition, it provides a general overview of the harmonization procedures outlining specific examples of how these procedures have been used.

4 HARMONIZATION PROCEDURES

The following steps are generally included in harmonization procedures: (a) Identifying data domains to be harmonized; (b) Compiling an inventory of relevant data across cohorts; (c) Defining project-specific research protocols (d) Harmonising specific variables under a common format; (e) Assessing the quality of harmonized data; (f) Preserving the data and associated metadata for future usage and replication.

4.1 Identifying data domains to be harmonized

The first step is to identify the different kinds of data that will be obtained by the members of the consortium, based on the objectives of each WP.

WP	Objective	Description
WP 3	O3.2: Estimate stability and change in genetic nurture effects throughout the life course	UCL will estimate genetic nurture and genetic transmission effects, using polygenic scores. UCL will investigate whether estimates of genetic nurture remain stable or change between childhood, adolescence, and adulthood using structural equation modelling.
WP 3	O3.3: Identify mechanisms underlying risk transmission and resilience factors	NIPH will investigate putative mediators including parenting, stressful life events, breastfeeding, mother-father relationship quality, parent-child relationship quality, and social support, that may act as parental resilience factors that attenuate offspring genetic risk.
WP 4	O4.2: Identify epigenetic markers of 'transmission load' as predictors of offspring outcomes	Epigenetic profiles of transmission load for mental illness will be identified in cord blood and tested for prediction of later offspring psychopathology. Classical statistics (e.g. regression models) and deep learning methods (MethylNet) will be used to model associations between parental mental illness and neonatal epigenetic patterns
WP 4	O4.3: Establish whether epigenetic patterns mediate parental mental illness effects on child mental health	New methods (DACT, MICS) will be used to quantify epigenome-wide mediation of parental illness on offspring mental outcomes. Mediation will also be analysed using poly-epigenetic scores of transmission load.
WP 5	O5.2: Identify neuroimaging markers of transmission load in childhood as predictors of	To (1) identify brain metrics at the age of 10 years and (2) test whether these brain metrics are predictive of

	mental health outcome in adolescence	later mental health problems in offspring at age 17 years.
WP 5	O5.4 Explore how genetic and environmental routes of transmission of parental mental health problems to offspring relate to brain features (from WP3) in childhood and adolescence	Triad genetic design from WP3 will be implemented to quantify (i) whether these polygenic scores associate with brain traits that are shared within-family, and (ii) how much of this is due to genetic transmission versus genetic nurture effects.
WP 7	O7.2: Translation of group-statistical patterns to the prospective prediction of mental health for individuals at high familial risk	RUMC will build normative models for the prospective prediction of mood and psychosis symptoms and diagnoses based on an optimal combination of genomic, epigenomic, brain, and environmental information.
WP 7	O7.3 Novel and flexible quantification of resilience to mental health problems and generation of new hypotheses of causal mechanisms underlying risk and resilience of mental health problems	These quantifications will be done using symptoms, overall clinical diagnosis as well as broader measures of functioning as outcome variables in GenR.

Table 2. FAMILY objectives involving human data analysis.

The core data domains that need to be harmonized based on these objectives have been identified. It should be noted that the offspring and parental cohorts will be harmonized separately due to the different nature inherent to these two populations and to the different objectives for each of them. The data domains to be harmonized are listed below:

- Genetics: UCL
- Epigenetics: EMC
- Neuroimaging: EMC
- Clinical data: FCRB, all sites
    - Socio-demographic information
    - Clinical and behavioural data (offspring and parental diagnoses, symptom severity, comorbidities, functioning, coping style)
    - Environmental factors (prenatal and postnatal, early life stress, parenting style)

4.2 Compiling an inventory of relevant data across cohorts

The next step will be to compile details about particular variables within the core data domains that need to be harmonized. A data inventory will be created and completed by all partners in order to collect information of relevant variables available across the FAMILY human cohorts. An example of the data inventory table to collect information across cohorts is provided in Table 3.

Table 3. Example of the data inventory table to collect information across cohorts.

			Parent Clinical Assessment				**Offspring clinical assessment**										
Cohort	**Partner**	**General population/ High risk**	**Parents Clinical assessment (yes/no)**	**Diagnostic tool**	**Socio-economic status**	**Global functioning**	**Offspring clinical assessment**	**Diagnostic tool**	**Global functioning**	**Early life stress**	**Parenting style**	**Coping**	**Prodromal psychotic symptoms**	**Bipolar prodromal symptoms**	**Breastfeeding**	**Self-esteem**	**Social support**
GenR/ORACLE	EMC	General population															
COPSYCH	RegionH	General population															
ALSAPAC		General population															
MoBa	NIPH	General population															
MCS	UCL	General population															
UK Biobank		General population															
ABCD		General population															
HCP		General population															
PNC		General population															
BRIDGE	EMC	High risk															
KBO	EMC	High risk															
MARIO	EMC	High risk															
BASYS	FCRB	High risk	yes	SCID	Hollings-Head Redlich	GAF	yes	KSADS/SCID	C-GAS/GAF	SLES	PBI		SIPS/SOPS	BPSS (at 8-year follow-up)			
BASYS	FIBHGM	High risk															
LG	CHUV	High risk															
VIA	RegionH	High risk															

SCID = Structured Clinical Interview for DSM; GAF = Global Assessment of Functioning scale; K-SADS = Kiddie-Schedule for Affective Disorders & Schizophrenia; C-GAS = Children's Global Assessment Scale; SLES = Stressful-Life-Experiences-Screening; PBI = Parental Bonding Instrument;

SIPS/SOPS = Structured Interview for Psychosis –Risk Syndromes (SIPS)/ Scale for the Assessment of Prodromal Symptoms (SOPS); BPSS = Bipolar Prodrome Symptom Scale.

4.3 Defining project-specific research protocols

4.3.1 Genetics

4.3.2 Epigenetics

4.3.3 Neuroimaging

Neuroimaging data will be converted from DICOM image format to nifti format using the dcm2niix converter (Li et al., 2016) and stored according to the Brain Imaging Data Structure (BIDS) specification (Gorgolewski et al., 2016). DICOM data stored within an XNAT (Herrick et al., 2016) instance will have immediate compatibility with FAMILY code for converting to nifti format and conforming to BIDS.

4.3.3.1 T1-weighted Structural MRI

T1-weighted structural MRI data will be processed through the FreeSurfer analysis suite (Fischl, 2012). Briefly, Freesurfer is an automated analysis software for structural MRI images which conducts non-brain tissue removal, B1 inhomogeneity correct (signal intensity normalization), tissue segmentation, tesselation of a white and pial surface, estimation of several morphological features (e.g., cortical thickness and folding), automatic anatomical labelling of several structures, and more. FreeSurfer version 6 and higher are eligible for use within FAMILY, with a preference for use of the latest version (v7.3.2). No contrasting beyond this on version will be made however, given the scale of datasets and the high correlation between data produced by version 6 and version 7. Version of FreeSurfer should be noted in the dataset and available for use as a covariate in statistical models. T1-weighted only inputs are preferred, though if T2 or FLAIR inputs were also used this will be accepted, though as with version must be noted and available as a covariate for analysis. Standard (default) options are preferred with the recon-all processing stream. If variation from the defaults has been implemented, this should be indicated and documented with the data. The "qcache" stream should be invoked after recon-all processing of streams 1, 2 and 3 have completed, and surface maps will be registered to "fsaverage" space with a single, FWHM smoothing kernel of 10mm. Region of interest (ROI) data will be tabulated with tools bundled with FreeSurfer, namely (e.g., aparcstats2table, https://surfer.nmr.mgh.harvard.edu/fswiki/aparcstats2table). R statistics code to convert the CSV outputs into R data structures and provide metadata attributes will be made available. A full Standard Operating Procedure for the FreeSurfer processing can be found in Appendix X and on the FAMILY Knowledge Base.

4.3.3.2 Diffusion MRI

4.3.3.3 Resting-state Functional MRI

The next steps will be developed once we have full information on data availability from all sites (4.3.4 Harmonizing specific variables under a common format: Clinical Data; 4.4. Assessing the quality of harmonized data; 4.5. Preserving the data for future usage and replication).

5 LITERATURE REFERENCES

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/J.NEUROIMAGE.2012.01.021

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data 2016 3:1*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.44

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). XNAT Central: Open sourcing imaging research data. *NeuroImage*, *124*(Pt B), 1093–1096. https://doi.org/10.1016/J.NEUROIMAGE.2015.06.076

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, *264*, 47–56. https://doi.org/10.1016/J.JNEUMETH.2016.03.001

6 APPENDICES

ACKNOWLEDGEMENT
Funded by the European Union, the Swiss State Secretariat for Education, Research and Innovation (SERI) and the UK Research and Innovation (UKRI) under the UK government’s Horizon Europe funding guarantee'. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, or the European Research Executive Agency (REA), the SERI or the UK Research and Innovation (UKRI). Neither the European Union nor the granting authorities can be held responsible for them.

FAMILY

“Running in the FAMILY – Understanding and predicting the intergenerational transmission of mental illness”

Grant Agreement number: 101057529

1st version of open science plan

Workpackage:	WP 2
Task:	Task 2.2
Due Date:	31st March 2023 (M6)
Actual Submission Date:	06th April 2023 (M7)
Project Dates:	Project Start Date: October 01, 2022 Project Duration: 60 months
Responsible partner:	Partner 01 EMC, 07 FCRB
Responsible author:	Lisanne van Houtum, Gisela Sugranyes
Email:	l.vanhoutum@erasmusmc.nl, GERNEST@clinic.cat
Contributors:	-

Project funded by the European Commission within HORIZON-HLTH-2021-STAYHLTH-01-02: Towards a molecular and neurobiological understanding of mental health and mental illness for the benefit of citizens and patients’		
Dissemination Level		
PU	Public — fully open (automatically posted online)	X
SEN	Sensitive — limited under the conditions of the Grant Agreement	

OPEN SCIENCE PLAN

FAMILY proposal:

1.2.7 Open science

Open access: EMC, as the Coordinator of FAMILY, will ensure the compliance with HORIZON rules regarding Open Access to scientific publications, by making all scientific publications generated in the project freely accessible. FAMILY will publish as much as possible of its work in peer-reviewed open-access journals (gold open access). FAMILY partners have included funds in their budgets to do so. Additionally, RePub, which is the institutional repository of the EMC (https://repub.eur.nl/), meets all the requirements established by the European Commission within the framework of open-access publishing (green open access). FAMILY researchers will have early full-text access to any new publication arising from FAMILY through the password-protected project intranet. Whenever possible, the full text of published articles or corresponding accepted manuscript will also be made available on the project website (portal; WP9) and spread out through concomitant updates on the relevant social media channels upon publication, particularly Twitter, LinkedIn and ResearchGate. All peer-reviewed publications will be deposited in Zenodo, the open-access archive funded by EC, the OpenAIRE project, CERN, PROSPERA, bioRxiv, medRxiv, and PsyArxiv, ensuring public availability of research materials including journal articles, conference proceedings, reports, deliverables, and presentations.

Open data and methods: Data from all participating cohorts in FAMILY will be made available to interested researchers and the scientific community at large by placing it in local or repositories hosted by the FAMILY consortium for use by others. For this, the Digital Research Environment (DRE; WP2,7) offers a promising solution allowing data use by external researchers while assuring full control by those partners who are responsible for the data. However, in studies involving human subjects, the principles of open science must be balanced with the need for data protection and privacy. WP2 will explicitly formulate the principles and procedures for maximizing open science in studies with human subjects in Standard Operating Procedures (SOPs) and will continuously monitor their implementation. SOPs will differ between cohorts, given that the informed consent that is provided by participants differ between cohorts, resulting in different procedures for data sharing or data accessibility. Where relevant, FAMILY's methodology will be made available on GitHub and on the FAMILY website. In publications, reports and presentations using the methods, researchers will be referred to FAMILY's GitHub project or website.

Research integrity & reproducibility of scientific results: Researchers in FAMILY will adhere to relevant standards for good research practices. A mentoring program where young researchers are linked with senior researchers in the consortium (but who are not directly involved in the young researcher's project) will be put in place as part of the training program, where issues related to research integrity can be raised and solved. Reproducibility of scientific findings will be facilitated in several ways. To facilitate sharing and long-term use of FAMILY's data, the following formats will be chosen: pdf, txt, csv, sql, dat (SPSS), RData, DICOM, NIfTI. All files will be marked with explicit dates (YYYY–MM–DD) and version numbers, where appropriate and provenance information will be documented in the Knowledge Base. FAMILY will use standardized variable names linked to meta data in a data dictionary. In cases where possible, meta data will be included inside of files (e.g. attributes within RData structures). Industry-standard data structures will be utilized for brain imaging data (Gorgolewski et al., 2016) and standardized processing pipelines will be applied to imaging and -omics data, in many cases within Singularity containers to ensure consistent and reproducible processing is applied uniformly to all data (Kurtzer et al., 2017). Further, the DRE platform allows for virtual machines to be generated and cloned, using precisely the same hardware and software infrastructure, ensuring all researchers work within the same environment, avoiding any platform-dependent biases. Version control and data provenance mechanisms (e.g. GitHub) will allow for consortium partners to track, archive, and publish their code in a transparent fashion. Data management and harmonization plans will be developed and documented on the web portal, and data provenance clearly established for all datasets.

Open science education and skills: FAMILY researchers will be offered access to education to develop the necessary skills and support to apply open science research routines and practices. In most institutions that participate in FAMILY, Open Science Communities are in place and/or Open Science Officers are employed (e.g. LIR, EMC, RUMC, UCL). FAMILY researchers will be encouraged to participate in activities that are organized locally. The EMC, Coordinator of FAMILY, is home to the Reproducible Interpretable Open Transparent (RIOT) Science Club Rotterdam, which originated from King's College London. The RIOT Science Club is a seminar series that raises awareness and provides training in Reproducible, Interpretable, Open & Transparent science practices. The initiative is entirely early career researcher-led and has now expanded beyond King's College to a growing number of sites (e.g. EMC, UCL) and is partnered with the UK Reproducibility Network. All presentation slides are stored on an Open Science Framework page: https://osf.io/8y7h2/, and recordings are uploaded to the RIOT Science Club YouTube channel. The FAMILY website will refer to these outlets and will stimulate its researchers to take full advantage of its content.

Citizen science: WP8 seeks active engagement of family members, patients, and mental health care professions, who can be reached via the European Federation of Associations of Families of people with Mental Illness (EUFAMI) and the European Society of Child and Adolescent Psychiatry (ESCAP). FAMILY will involve EUFAMI and ESCAP and they will support FAMILY with dissemination to their established communication channels. EUFAMI will reach 32 family organizations in 21 countries throughout Europe and ESCAP has 34 national member societies from 33 European countries. They will greatly facilitate direct access to relevant stakeholders for their contribution to WP8 as well as uptake of new knowledge by the clinical, scientific, policy making communities.

WP2 pProvides support and resources to facilitate open science practices across all its tasks. (Task 2.3)

Task 2.2 Construct a data management, harmonization, and open science plan (FCRB, EMC) - (M1-M60)

FCRB and EMC will create a Data Management Plan based on FAIR principles. FCRB will create a web-based Knowledge Base which will include the plan and all accompanying SOPs, including a detailed and searchable data dictionary of all raw and derived data. In addition to general management and harmonization of data, the plan will also include important information regarding GDPR, and protocols for collecting and collating all necessary documentation related to data from partners, including data use/transfer and confidentiality agreements. Further, WP2 will work closely with WP8 in ensuring ethical guidelines for use and reuse of data are well-documented in the Knowledge Base. Working closely with WP9, all documentation necessary for facilitating open publication of work from the project will be integrated into the web portal's Knowledge Base. Further, a GitHub repository will be created for the consortium and linked to the web portal, so that code from each WP can be openly shared with detailed version histories available.

International collaborations:

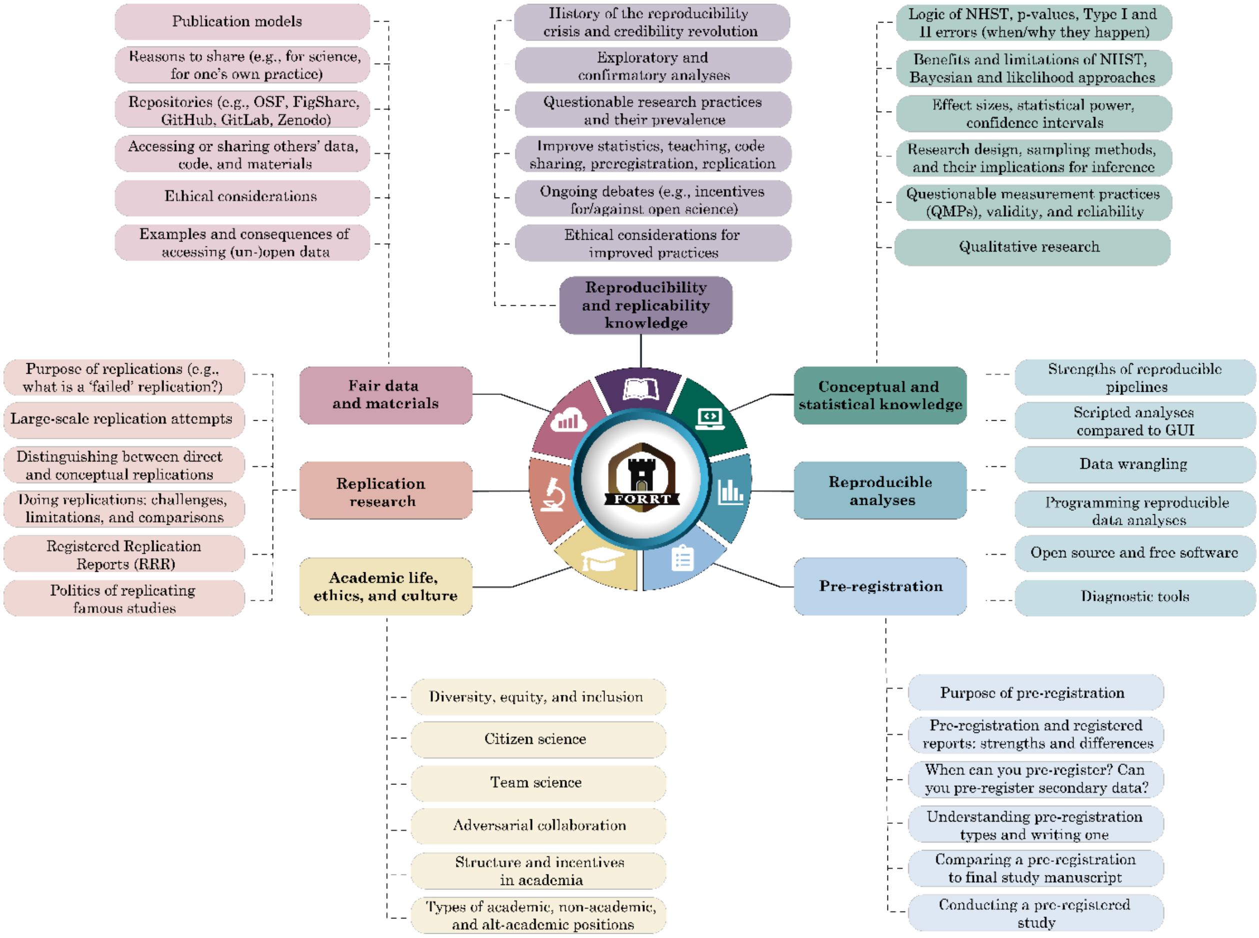
ROSiE UL (WP leader: Mezinska): Developing tools to ensure ethics and research integrity in open science. The tools will be applied for responsible practice of open science within the FAMILY.

Useful links:

[OSF | Open Science Initiative in Psychology @LMU]() This OSF project collects documents, presentations, etc. from the Open Science Initiative at the psychology department of the Ludwig-Maximilians-Universität München.

[OSF | Open Science Crash Course (everything in 5 hours)]() Open Science Workshop Materials of the LMU Open Science Center

https://forrt.org/clusters/


FORRT
Reproducibility and replicability knowledge
History of the reproducibility crisis and credibility revolution
Exploratory and confirmatory analyses
Questionable research practices and their prevalence
Improve statistics, teaching, code sharing, preregistration, replication
Ongoing debates (e.g., incentives for/against open science)
Ethical considerations for improved practices
Fair data and materials
Publication models
Reasons to share (e.g., for science, for one's own practice)
Repositories (e.g., OSF, FigShare, GitHub, GitLab, Zenodo)
Accessing or sharing others' data, code, and materials
Ethical considerations
Examples and consequences of accessing (un-)open data
Conceptual and statistical knowledge
Logic of NHST, p-values, Type I and II errors (when/why they happen)
Benefits and limitations of NHST, Bayesian and likelihood approaches
Effect sizes, statistical power, confidence intervals
Research design, sampling methods, and their implications for inference
Questionable measurement practices (QMPs), validity, and reliability
Qualitative research
Replication research
Purpose of replications (e.g., what is a 'failed' replication?)
Large-scale replication attempts
Distinguishing between direct and conceptual replications
Doing replications: challenges, limitations, and comparisons
Registered Replication Reports (RRR)
Politics of replicating famous studies
Reproducible analyses
Strengths of reproducible pipelines
Scripted analyses compared to GUI
Data wrangling
Programming reproducible data analyses
Open source and free software
Diagnostic tools
Academic life, ethics, and culture
Diversity, equity, and inclusion
Citizen science
Team science
Adversarial collaboration
Structure and incentives in academia
Types of academic, non-academic, and alt-academic positions
Pre-registration
Purpose of pre-registration
Pre-registration and registered reports: strengths and differences
When can you pre-register? Can you pre-register secondary data?
Understanding pre-registration types and writing one
Comparing a pre-registration to final study manuscript
Conducting a pre-registered study