

Common Infrastructure for National Cohorts in Europe, Canada, and Africa - CINECA -

Deliverable D7.4 Data Management Plan, v1.0

WP7 – Ethical and legal governance framework for transnational data-sharing

Lead Beneficiary:	European Molecular Biology Laboratory
WP Leaders:	Emmanuelle Rial-Sebbag (INSERM) Michaela Th. Mayrhofer (BBMRI-ERIC)
Contributing partners:	EMBL-EBI, HES-SO, UCT, SFU
Contractual delivery date:	30 June 2019
Actual delivery date:	27 November 2019
Authors of this deliverable:	Giselle Kerry, Leslie Glass, Thomas Keane, Patrick Ruch, Nicola Mulder, Fiona Brinkman
Reviewed by:	Michaela Th. Mayrhofer, Dylan Spalding, Éloïse Gennet
Approved by:	Michaela Th. Mayrhofer, Emmanuelle Rial-Sebbag
Dissemination Level:	Public
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	48 months

Table of contents:

1. Executive Summary	4
2. Data description and collection or re-use of existing data	4
2.1 How will new data be collected or produced and/or how will existing data be reused?	4
2.2 What type of data (for example the kinds, formats, and volumes) will be utilised?	4
2.3 What data will be exchanged between different partners?	5
2.4 How will these data be accessed by the researchers?	5
3. Data processing, documentation and quality control	5
3.1 What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?	5
3.2 What data quality control measures will be used?	7
3.3 Who is responsible for data management?	7
3.4 What systems and software will be used to process the data?	7
3.5 How will the project deal with data changes during the project (eg, versioning issues)?	8
3.6 What financial, time or other resources will be dedicated to data management and FAIR?	8
4. Storage and backup during the research process	9
4.1 How will data security and protection of sensitive data be taken care of during the research?	9
5. Legal and ethical requirements, codes of conduct	9
5.1 If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?	9
5.2 How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?	9
5.3 How will possible ethical issues be taken into account, and codes of conduct followed?	10
5.4 How will data security and protection of sensitive data be taken care of during the research?	10
6. Fair data re-use and long-term preservation beyond the project	11
6.1 Findability: how will you promote that your data can be found by interested users?	11
6.2 Accessibility: what methods or software tools will be needed to access and use the data?	11
6.3 Interoperability: how is interoperability of data and software promoted?	11
6.4 Reusability: what metadata will accompany input, intermediate and result data?	12
6.5 How will the application of a unique and persistent identifier (such as Digital Object Identifier (DOI)) to each data set be ensured?	12
7. Data management responsibilities and resources	12
7.1 Who (for example role, position, and institution) will be responsible for data management (ie, the data steward)?	12
7.2 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?	12
8. References/Resources	12
9. Abbreviations	13

10. Delivery and schedule	14
11. Appendices	14
Appendix 1: List of CINECA cohorts and the process for accessing secondary use data identified by CINECA.	14
Appendix 2: CINECA Work Packages	15
Appendix 3: CINECA cohorts and types of data available.	16

1. Executive Summary

The overarching purpose of CINECA is to achieve federated human data interoperability between 10 existing cohorts from Canada, Europe and Africa which represent >1.4 million individuals. This will enable population scale genomic and biomolecular data access across international borders, accelerating research and improving the health of individuals resident across continents. This project will not generate novel data from human data, rather it relies on integrating existing resources which hold or store data to deliver new knowledge and innovation. All data access is determined by the existing data access committees (DACs) and the respective data processes for each dataset (Appendix 1 explains the process for accessing secondary use data for each cohort). The informed consent and ethics approvals for CINECA cohorts were documented in D9.3 and D9.4 and will be respected by this data management plan, which will be an integral part of the consortium's Governance Framework. Given the international nature of this project, the data management plan is a component of Work Package 7 (WP7 Ethical and legal governance framework for transnational data-sharing) (All WPs are listed in Appendix 2). We note that all partners in the project have relevant national and/or international experience in acquiring, storing, analysing and sharing high complexity datasets and in the implementation of the FAIR principles for these resources.

The Data Management Plan (DMP) was developed based on the core requirements for DMPs as described by The Science Europe Practical Guide to the International Alignment of Research Data Management (<https://www.scienceeurope.org>). The DMP is a living document, expected to be updated during the lifetime of the project.

2. Data description and collection or re-use of existing data

2.1 How will new data be collected or produced and/or how will existing data be reused?

CINECA is concerned with the secondary use of cohort data and metadata. No clinical, tracking or genetic data related to individuals or groups of individuals will be collected de novo by the participants for the purposes of executing this project. All data to be used has been collected by existing cohorts, which have in place existing ethical governance consistent with the legal framework for the country of origin. This project will not attempt to circumvent these and will adhere to these processes. All personally identifiable data is pseudonymised at the originating cohort prior to being made available for secondary use by CINECA partners.

2.2 What type of data (for example the kinds, formats, and volumes) will be utilised?

Data across the CINECA cohorts varies widely due to the differing types of biomaterials collected from participants (e.g. blood, DNA, tissue samples), study design (eg, longitudinal,

disease-specific), lifestyle information provided by participant questionnaires and molecular measurements to assess phenotypes. A list of initial cohorts is included in Appendix 1, CINECA will only use data that is in a digital format and will include data from:

- Surveys
- Questionnaires and interviews
- Registry data (observations)
- Clinical measurements
- Medical records
- Electronic health records
- Administrative records
- Environmental data collection
- Biological samples
- Genetic and other OMIC data
- Intermediate statistics for meta-analysis

2.3 What data will be exchanged between different partners?

CINECA consortium members will follow the cohorts existing procedures to request access to cohort data for which full consent for sharing has been obtained from participants. The precise data to be exchanged will vary depending on the application, e.g. federated GWAS will use phenotype and molecular data from cohorts, federated genotyping will use molecular data such as exome or whole-genome sequencing, eQTL analysis will use gene expression levels

2.4 How will these data be accessed by the researchers?

The data will be accessed by researchers initially using the existing cohort access interfaces (e.g. secure bulk data copying/transfer), and as the work of WP1 (Data discovery), WP2 (Interoperable Authentication and Authorisation), WP3 (Harmonised Metadata) is completed, based on Global Alliance for Genomics and Health (GA4GH¹) standards, we expect WP4 (Federated Joint Cohort Analysis) and WP5 (Healthcare Interoperability and Clinical Applications) to use these standards for discovery and access to cohort data.

3. Data processing, documentation and quality control

3.1 What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany data?

The metadata and documentation that will accompany cohort data will vary by cohort according to what data and meta data was collected and made available to researchers

¹ <https://www.ga4gh.org/>

(Appendix 3). For example, the European Genome-phenome Archive² (EGA) contains data from >2,000 human genetic and phenotype research studies across a variety of diseases and data types. The metadata provided to authorised users varies depending on individual studies, while many studies provide participant metadata via commonly used ontologies, such as the Human Phenotype Ontology (HPO) or Experimental Factor Ontology (EFO), such as H3Africa. EGA collects comprehensive metadata on experimental methodology, such as sequencing technologies or analysis pipelines, and provides comprehensive documentation to users via its website and as downloadable xml files. This approach to comprehensive documentation is replicated for most of the other cohorts participating in CINECA.

H3Africa is attempting to standardize its metadata collection by developing standardized case report forms (CRFs) for core phenotypes with additional disease-specific modules. These have been converted into REDCap data dictionaries, which many projects are using. H3Africa has also developed guidelines for creating and completing the CRFs³. Since H3Africa is developing a metadata catalogue for H3Africa data and biospecimens, they have also created a set of metadata elements that will be uploaded into the catalogue for each H3Africa study. For the data going to the EGA we provide metadata templates to ensure consistency in reporting and wherever possible we annotate to ontologies such as disease ontology, human phenotype ontology and experimental factor ontology.

CoLaus/PsyCoLaus comes with different types of metadata. Before access is granted a textual description of the cohort can be found online, to be referenced by CINECA via links to the cohort website⁴. Regarding structured metadata, there is DATS file available, which contains information about the therapeutic areas of the cohorts and the healthcare procedures (diagnoses are codes via ICD-10, prescription via WHO-ATC), and the type of the population (demographics). Finally, once access is granted by the DAC, a detailed code book is shared, so that the variables of interest for the researchers can be selected before signing the Data Transfer Agreement. In addition to prescription and diagnosis (main and comorbidities), the variables also include LOINC lab results.

With respect to the CHILD cohort (Canadian Healthy Infant Longitudinal Development Study⁵) metadata is extensive (over 200 questionnaires, 47 million data points, for just one component of the study) and wide ranging (home environment, socioeconomic status/demographics, food/diet, diseases and symptoms, medications, pets, stress/mental health, work/school environment, pollution and other exposures, stool/breast and milk/dust microbiome, genome, epigenome, transcriptome, metabolome, chemical analytics, immune function, lung function and other organ-based tests).

² <https://www.ebi.ac.uk/ega/home>

³ <https://h3abionet.org/data-standards/datastds>

⁴ <https://www.colaus-psycolaus.ch/colaus/>

⁵ <https://childstudy.ca/>

Descriptions of the data methodology are made available in documents and publications, along with the original data, cleaned data, derived variables (with pseudocode documenting how they are calculated), and assigned ontology terms using a collection of well-developed ontologies associated with the Open Biological and Biomedical Ontology (OBO) Foundry⁶ and other accepted ontologies. Two examples reflecting different levels of complexity: 1. CHEBI (Chemical Entities of Biological Interest⁷) ontology terms are assigned for medications, with ATCC (American Type Culture Collection) codes. 2. A collection of five ontologies are used to assign terms for "diseases and symptoms" data: Symptom Ontology, Disease Ontology, Ontology of Adverse Events, Human Phenotype Ontology, and National Cancer Institute Thesaurus (in an initial curation round we found these five ontologies to be the most useful for organizing these data). In some cases new ontologies are being created, in particular a new asthma ontology is currently being designed, with plans for wider consultation and then submission of the proposed ontology for consideration to the OBO Foundry. In addition to ontologies, summary statistics are provided for each questionnaire question (and for clinical or laboratory data), viewable through a web interface after registration for access. Additional info through a queryable web interface is provided (for example, all data types/questions associated with asthma), and an overview of CHILD data is additionally made available through the Maelstrom resource⁸.

3.2 What data quality control measures will be used?

The primary cohort data quality controls are expected to be carried out by the individual cohort owners or data producers. In addition, the federated research and clinical applications in WP5+6 will carry out specific QC on cohort datasets to ensure suitability and sufficient quality for the specific application, examples include Hardy Weinberg Equilibrium (HWE) for population molecular data such as genotyping, format validation to ensure correct functioning with existing tools, or gene expression level consistency for eQTL analysis.

3.3 Who is responsible for data management?

Cohort data management is the responsibility of the cohort owners. Once access has been granted to a CINECA Principal Investigator, management of cohort data is regulated by the Data Access Agreements (DAA) signed between the cohort owner, the CINECA PI, and the signing official at their institute.

3.4 What systems and software will be used to process the data?

In CINECA, data processing only takes place in WP4+5 (Research and clinical federated applications). Data will be processed by a variety of tools depending on the research or

⁶ <http://www.obofoundry.org/>

⁷ <https://www.ebi.ac.uk/chebi/>

⁸ <https://www.maelstrom-research.org/>

clinical application in question. The precise list of software will be provided as WP4+5 carry out their work during the project.

3.5 How will the project deal with data changes during the project (eg, versioning issues)?

Versioning of cohort datasets is controlled by the respective cohort owner. Policies for notifying authorised users of dataset updates are documented through the specific DAAs or documented cohort standard operation procedures (SOPs). For example, the EGA tracks all changes to a dataset object, including both addition and withdrawal of data. Currently the EGA does not make the versions of a specific object public allowing a particular version of a dataset to be referenced by a unique identifier, hence EGA recommends that when a dataset is modified it is withdrawn and replaced by the new modified version with its own accession. The EGA is in the process of introducing version numbers to datasets and associated objects, improving the transparency of dataset changes.

For H3Africa data are only made available through public archives such as EGA when the final version is available. Within datasets the use of version numbers is encouraged if changes are expected. Once H3Africa is in EGA it will follow the policy listed above for versioning of datasets. For CoLaus/PsyCoLaus, H3Africa does not expect any change during CINECA. The study has been completed in 2018. New follow up recruitments may occur in the future, but this is not currently planned. In case of change in the source cohort files, H3Africa should be notified by the cohort owner, which may trigger a new version of the data. It will replace the obsolete version by the new version.

Regarding CHILD, the main cohort curation database contains original data, changed data (all tracked with a queryable log) and ontologies (all changes also tracked). An additional NoSQL extended database contains additional large omics/experimental/environmental data, with raw data stored on a file server, and processed data using multiple approaches (again, all methodology for derived variables tracked with pseudocode to document them). The data is housed in one primary location with real-time version tracking, with copies in two other locations across Canada (planned nightly syncing not yet implemented). CHILD are able to take advantage of SFU's high performance compute infrastructure which meets all security requirements for clinical data hosting, and includes robust climate control, full backup (including full power backup), and has been this year awarded \$26 million in further infrastructure – almost half of the total \$55 million in the initial round of Digital Research Infrastructure allocation spread across the 5 main data centres in Canada. SFU is providing in-kind infrastructure and administrative support for CHILDb (the database for the CHILD cohort) and housing of the large omics datasets.

3.6 What financial, time or other resources will be dedicated to data management and FAIR?

The fundamental goal of the CINECA project is to increase the level of FAIRness across the participating cohorts by engaging with the cohort owners to increase discoverability of molecular and phenotype data and metadata (WP1), create an interoperability network for data access by linking the cohorts via the ELIXIR and CanDIG AAls (WP2), and harmonise the cohort metadata and data dictionaries (WP3). Almost 50% of direct costs of the CINECA funding is dedicated to these efforts.

4. Storage and backup during the research process

4.1 How will data security and protection of sensitive data be taken care of during the research?

All cohort data access must adhere to the terms of a DAA under which the submitting institutions requirements must be agreed to before data access is granted. By managing the data in this way under the CINECA project, we can be reassured that the data is maintained at the correct level of security and only shared in accordance with the terms of the original participant consent and local laws.

5. Legal and ethical requirements, codes of conduct

5.1 If personal data are processed, how will compliance with legislation on personal data and on data security be ensured?

This project will not generate novel data from human data, rather it relies on integrating existing resources which hold, store or store data to deliver new knowledge and innovation. Intensified sample and data-sharing is a priority indicator for the success of for large scale cohort and biobanks' activities. Given the new methodologies used in health research, we face a desire and need to increase the reuse of human biological material. However, researchers face practical challenges to meet the legal and ethical requirements accordingly and in good time. From the legal point of view the two elements composing a biobank and/or cohort (namely samples and data) are usually covered by two different pieces of legislation in almost all the European countries as well as in Canada. Human biological samples are falling under bioethical principles such as non-commercialisation or human dignity whereas data are regulated under technical standards and protection of individual privacy, for example the General Data Protection Regulation (GDPR) in Europe and The Personal Information Protection and Electronic Documents Act in Canada (PIPEDA). All data access will be governed by a DAA between the cohort owner, the CINECA PI, and the PI's institute. All DAAs are required to be compliant with local applicable laws and regulations, such as GDPR in Europe, and PIPEDA in Canada.

5.2 How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

The participating cohorts/biobanks remain the owner of the cohort all data shared during the project, as governed by the DAAs. All software will be published as open source. A Consortium Agreement (CA) based on the DESCA Horizon 2020 model has been negotiated between all partners as a basis for the management of both the project and any resulting intellectual property rights (IPR). The CA specifies all arrangements, rights, and responsibilities of participants regarding the IPR already available prior to the start of the project and needed for the execution of the project.

5.3 How will possible ethical issues be considered, and codes of conduct followed?

The CINECA project is designed to develop common infrastructures for national cohorts in Europe, Canada, and Africa. CINECA's objectives will be compatible with EU Regulation No 536/2014 (2001/20/EC Directive), and the latest version of the Declaration of Helsinki, which since 2016 is complemented by the Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks, and will follow all relevant principles. Human biological data are regulated under technical standards and protection of individual privacy (GDPR in the EU and PIPEDA in Canada).

The CINECA project has a detailed WP (WP7) dedicated to defining the ethical and legal governance framework for transnational data-sharing. We have defined tasks and deliverables dedicated to identifying and cataloguing ethical and legal gaps across the Canadian, European and African cohorts, and providing recommendations for a Governance Framework. This will provide the project with accurate and well-grounded ethical and legal recommendations for the implementation of an appropriate data-sharing flow between European, Canadian, and African cohorts to intensify collaboration and promote reuse of samples and data for scientific knowledge gain. Ethics in the CINECA project will be reviewed extensively by the Ethics Management Group, membership of which includes the Data Protection Officers (DPOs) from all participants. Ethics reporting will occur at regular intervals in the project (M18, M36, M48). An Independent European Ethics Expert will review Ethical reports for CINECA, ensuring compliance with European standards as a minimum across the project.

All cohorts have their own responsibility for local ethical and participant consent. Only pseudonymised data will be shared within the consortium. All needed ethical approvals from the institution and participant for relevant data collection, use and sharing are the responsibility of the partners. The CINECA coordination team has copies of these approvals (WP 9 deliverables, submitted June 2019). All ethical approvals of partners and their cohorts are available upon request. The ownership and primary responsibility of data are always with the institution, which runs its own cohort data collection.

5.4 How will data security and protection of sensitive data be taken care of during the research?

Cohort data that is shared between CINECA partners will be governed by a DAA between the cohort, the CINECA PI, and the PI's institute. The DAAs outline the requirements for data security and protection of all sensitive data transferred from the cohort to the CINECA partner. The CINECA partner institutes have established local policies for handling sensitive data governed by DAAs and will comply with DAA requirements.

6. Fair data re-use and long-term preservation beyond the project

6.1 Findability: how will you promote that your data can be found by interested users?

The goal of WP1 in CINECA is to enable federated discovery of cohort data by a variety of methods, e.g. phenotype, genotype, and data use. WP1 will achieve this by:

- Building a federated system capable of executing cross-dataset and cross-institution queries that enables variables to be updated at source and longitudinal data to be analysed;
- Build on and contribute back to the GA4GH standards for responsible genomics data sharing in the Discovery Work Stream (DWS);
- Leverage and extend a set of technical standards like Beacon (Fiume, *et al*) and benchmark and implement these standards, and related tools for enabling rapid, flexible, secure, and federated discovery;
- Enable the querying of the genomic, phenotypic data, and associated metadata across datasets made available by the partners.

6.2 Accessibility: what methods or software tools will be needed to access and use the data?

CINECA will contribute to the development and implementation of data access interfaces defined by GA4GH. The GA4GH has membership from a variety of genome data use-cases such as industry, academia, medical centres, research institutes, national genome initiatives. Alignment of CINECA with GA4GH will ensure that access and interoperability can be achieved using community tools, e.g. the *htsget* API for streaming and accessing genomic data has been implemented by a variety of community tools and cloud platforms (Kelleher *et al.*), and the Researcher Identity and Authorisation and Authentication Infrastructure standards developed by GA4GH.

6.3 Interoperability: how is interoperability of data and software promoted?

CINECA will contribute to the development of the GA4GH standards and implement those interfaces at the participating cohorts. This will increase the interoperability of the cohort

data and metadata. Examples include the Data Use Ontology (DUO) which is a standard ontology for defining what type of research can be carried out on cohort data. The mapping of cohort data use conditions to DUO will enable faster discovery of cohort data suitable for specific research.

6.4 Reusability: what metadata will accompany input, intermediate and result data?

The goal of CINECA is to increase the reusability of cohort data by making the cohort data interoperable. WP1-3 will contribute to this goal.

6.5 How will the application of a unique and persistent identifier (such as Digital Object Identifier (DOI)) to each data set be ensured?

Individual participating cohorts in CINECA have existing schemes of assigning unique and persistent identifiers. For example, the EGA assigns unique identifiers to objects submitted, and objects are registered at identifiers.org under the EGA namespace. These identifiers are permanent and unique and are exposed through a Metadata API to encourage automatic indexing and discovery.

7. Data management responsibilities and resources

7.1 Who (for example role, position, and institution) will be responsible for data management (ie, the data steward)?

The DMP will be reviewed and updated annually following the CINECA AGM by the CINECA Project Manager and Coordinator, and more frequently should the need arise.

7.2 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

The goal of CINECA WP1-3 is to increase the FAIRness of the participating cohorts. Approximately 47% of CINECA direct costs is dedicated to these efforts.

8. References/Resources

DESCA Horizon 2020 model <http://www.desca-2020.eu/>

FAIR Data Management at a glance: issues to cover in your Horizon 2020 DMP. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf#page=10

Fiume M, *et al.* (2019) Federated discovery and sharing of genomic data using Beacons. *Nature Biotechnology*, 37:220–224. doi: 10.1038/s41587-019-0046-x.

Kelleher J, *et al.* (2019) htsget: a protocol for securely streaming genomic data. *Bioinformatics*, 35:119-121. <https://doi.org/10.1093/bioinformatics/bty492>.

Regulation EU No 536/2014 (2001/20/EC Directive)

SCIENCE EUROPE PRACTICAL GUIDE TO THE INTERNATIONAL ALIGNMENT OF RESEARCH DATA MANAGEMENT. 2 Nov 2018 - November 2018. 'Practical Guide to the International Alignment of Research Data Management': D/2018/13.324/4. Author: Science Europe.

World Medical Association. (2001). World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organization*, 79 (4), 373 - 374. World Health Organization. <https://apps.who.int/iris/handle/10665/268312>

World Medical Association. (2017). WMA Declaration of Taipei on Ethical Considerations regarding Health Databases and Biobanks. <https://www.wma.net/policies-post/wma-declaration-of-taipei-on-ethical-considerations-regarding-health-databases-and-biobanks/>

9. Abbreviations

AAI	Authentication and Authorisation Infrastructure
API	Application Program Interface
CA	Consortium Agreement
CHEBI	Chemical Entities of Biological Interest
CRF	Case Report Form
DAA	Data Access Agreement
DAC	Data Access Committee
DATS	DAta Tag Suite
DWS	Discovery Work Stream
DMP	Data Management Plan
DOI	Digital Object Identifier
DUO	Data Use Ontology
EFO	Experimental Factor Ontology
EGA	European Genome-phenome Archive
eQTL	expression Quantitative Trait Loci
FAIR	Findable, Accessible, Interoperable, Re-usable
GA4GH	Global Alliance for Genomics and Health
GDPR	General Data Protection Regulation

GWAS	Genome-Wide Association Studies
H3Africa	Human Heredity and Health in Africa
HPO	Human Phenotype Ontology
HWE	Hardy Weinberg Equilibrium
IPR	Intellectual Property Rights
NoSQL	Not only Structured Query Language
OBO	Open Biological and Biomedical Ontology
PI	Principal Investigator
PIPEDA	Personal Information Protection and Electronic Documents Act
SOP	Standard Operating Procedure
QC	Quality Control
WP	Work Package

10. Delivery and schedule

The delivery is delayed: Yes

The deliverable has been delayed in part due to miscommunication between the Project Coordinator and the Project Officer. The request for an extension to the due date had been informally approved to enable alignment of DMPs with the EUCAN-Connect project. Subsequently this was rejected in the amended Grant Agreement and the DMP deliverable was submitted immediately afterwards.

11. Appendices

Appendix 1: List of CINECA cohorts and the process for accessing secondary use data identified by CINECA.

Cohort Name	Access Process
CHILD, Canada	Contact National Coordinating Centre by email
CARTaGENE, Canada	Electronic access application via https://sdas.cartagene.qc.ca/
Canadian Longitudinal Study of Aging, Canada	Contact study coordinator by email
H3Africa, Pan country governance model	Request by emailed standard form for data hosted in and accessed from EGA. Access is controlled by H3Africa Data and

	Biospecimen Access Committee, governed by the consortium's Data Sharing, Access and Release policy
UK Biobank, UK	Application by UK Biobank electronic access system
European Genome-phenome Archive, multiple countries of origin	Application to individual DACs linked individual EGA datasets
BIOS, Netherlands	Application to individual DAC linked to EGA's datasets or access via NL centralised compute infrastructure
Estonian Biobank, Estonia	Application form submission to Ethics Review Committee University of Tartu
CoLaus, Switzerland	Contact study coordinator by email
PsyCoLaus, Switzerland	Contact study coordinator by email

Appendix 2: CINECA Work Packages

- WP1 Federated Data Discovery and Querying
- WP2 Interoperable Authentication and Authorisation Infrastructure
- WP3 Cohort Level Meta Data Representation
- WP4 Federated Joint Cohort Analysis
- WP5 Healthcare Interoperability and Clinical Applications
- WP6 Outreach, training and dissemination
- WP7 Ethical and legal governance framework for transnational data-sharing
- WP8 Project Management and coordination
- WP9 Ethics requirements

Appendix 3: CINECA cohorts and types of data available.

Cohort/ Resource name	Number of participants	Location	Longitudinal	Diseases	Gender	WGS	WES	RNASeq	Epigenetics	Genotyping
CHILD	3.5k	Canada	X	Population based developmental health and disease	M & F	X		X	X	X
CARTaGENE	43.0k	Canada	X	Population based cohort	M & F	X		X		X
CLSA	50.0k	Canada	X	Population based cohort	M & F					X
H3Africa	75.0k	South Africa		Multiple communicable and non-communicable diseases in multiple African countries	M & F	X	X			X
B IOS	4.0k	Netherlands		Population based cohort	M & F	X		X	X	X
Estonian Biobank	51.0k	Estonia	X	Population based cohort	M & F	X	X	X	X	X
Colaus	6.1k	Switzerland	X	Cardiovascular diseases	M & F		X			
PsyColaus	3.6k	Switzerland	X	Mental disorders	M & F		X			
EGA	700.0k	UK + Spain		Multiple diseases and healthy cohorts	M & F	X	X	X	X	X
UK Biobank	500.0k	UK	X	Population cohort and disease; cancer, heart disease, stroke, diabetes, arthritis, osteoporosis, eye disorder, depression and form of dementia	M & F	X	X			X