An EU-Canada joint infrastructure
for next-generation multi-Study Heart research

Deliverable D3.1:

Data Management Plan

Reference	D3.1_euCanSHare_CRG_31052019
Lead Beneficiary	CRG
Author(s)	Claudia Vasallo
Dissemination level	Public
Type	ORDP: Open Research Data Pilot
Official Delivery Date	31/05/2019
Date of validation by the WP Leader	31/05/2019
Date of validation by the Coordinator	31/05/2019
Signature of the Coordinator	

euCanSHare is funded by the European Union's H2020 Framework under Grant Agreement 825903.

Version Log

Issue Date	Version	Involved	Comments
17/05/2019	v.1.01	Claudia Vasallo (CRG)	1st draft
20/05/2019	v.1.02	Audald Lloret-Villas (CRG), Katharina Heil (UPF)	1st comments and additions
28/05/2019	v.2.01	Babita Singh (CRG), Audald Lloret-Villas (CRG)	Updates and Revision
29/05/2019	v2.02	Jordi Rambla	Revision
30/05/2019		Karim Lekadir (UPF), Katharina Heil (UPF)	Comments and review
30/05/2019	v3	Claudia Vasallo (CRG)	Integration of comments and update of deliverable
31/05/2019		Josep Lluis Gelpi (BSC)	Minor comments and suggestions
31/05/2019		Claudia Vasallo (CRG)	Final Review and finalization
31/05/2019		Aad van der Lugt (EMC)	Review through Work Package Leader
31/05/2019	final	Karim Lekadir (UPF), Katharina Heil (UPF)	Revised and corrected final version. Final touches and submission

Executive Summary

This Data Management Plan addresses the purpose and description of data handled within the euCanSHare project and the model for data handling during and after the project, including provisions for data collection, secure long-term storage, integration and interoperability, accessibility and exploitation, in compliance with the principles for findable, accessible, interoperable and reusable research data.

Table of Contents

List of tables

List of figures

Acronyms

DMP: Data Management Plan

FAIR: findable, accessible, interoperable, reusable

WP: work package

DAC: Data Access Committee

EGA: European Genome-phenome Archive

BBMRI: BioMolecular resources Research Infrastructure

1 Data summary

Data are a core component of euCanSHare as the project will centralise cardiovascular research data as a means to improve data discoverability, foster the development of technologies and analysis capacities and lead to cutting-edge collaborative research in the domain of cardiovascular medicine.

1.1 Purpose of data collection

EuCanSHare project will develop infrastructure and technology to centralise previously collected cardiology research data from Europe and Canada and to foster the collection and incorporation of new relevant data into a comprehensive cross-border and FAIR[1] (findable, accessible, interoperable and reusable) cardiovascular research network, thereby enhancing cardiology data sharing, discoverability and exploitation.

The data and the expertise on data collection, data management and metadata generation from pre-existing and future cardiology initiatives will be leveraged during the project to build a highly comprehensive, robust, secure and scalable cardiovascular data platform and to develop specialised data management and analysis tools for research in cardiology.

As personalised medicine approaches are urgently needed in cardiovascular research to improve risk assessment and prevention, early diagnosis, as well as for treatment personalization and drug development[2], data and data analysis tools centralised within euCanSHare are of great potential interest for both cardiology researchers and medical staff.

1.2 Description of data

EuCanSHare will include diverse data types from pre-existing and future prospective studies monitoring a wide range of baseline and follow-up measures as well as cardiovascular-relevant clinical outcomes.

In its initial phase, euCanSHare will incorporate data from pre-existing cardiovascular initiatives, such as the MORGAM/BiomarCaRE Project [3,4] and The Canadian Alliance for Healthy Hearts and Minds (CAHHM)[5], aimed at assessing classic and genetic cardiovascular risk factors, as well as cardiology-relevant data from broad scoped health initiatives such as the UK Biobank [6], the Study of Health in Pomerania (SHIP) [7] and the Hamburg City Study [8]. A total of over 35 cohorts from these existing initiatives is expected to be incorporated during this phase, adding up to over 1,000,000 records.

[1] FAIR principles for data stewardship. *Nat Genet* Nature Publishing Group; 2016;48(4):343–343.

[2] Mullard A. 2012 FDA drug approvals. *Nature reviews. Drug discovery* 2013;12(2):87–90.

[3] Blankenberg S, et al., MORGAM Project. Contribution of 30 biomarkers to 10-year cardiovascular risk estimation in 2 population cohorts: the MONICA, risk, genetics, archiving, and monograph (MORGAM) biomarker project. *Circulation* Lippincott Williams & Wilkins; 2010;121(22):2388–2397.

[4] Zeller T, et al., BiomarCaRE: rationale and design of the European BiomarCaRE project including 300,000 participants from 13 European countries. *European Journal of Epidemiology* 2014;29(10):777–790.

[5] Anand SS, et al., Rationale, design, and methods for Canadian alliance for healthy hearts and minds cohort study (CAHHM) - a Pan Canadian cohort study. *BMC Public Health* BioMed Central; 2016;16(1):650.

[6] Littlejohns TJ, et al., UK Biobank: opportunities for cardiovascular research. Eur Heart J 2019;40(14):1158–1166.

[7] Völzke H. Study of Health in Pomerania (SHIP). Concept, design and selected results. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz Springer-Verlag; 2012;55(6-7):790–794.

[8] Bohnen S, et al., Cardiovascular magnetic resonance imaging in the prospective, population-based, Hamburg City Health cohort study: objectives and design. J Cardiovasc Magn Reson BioMed Central; 2018;20(1):68.

Table 1 and Appendix Table 5 show a description of the 24 first cohorts being incorporated to euCanSHare within the first months of the project. As new cohorts are incorporated to the project the DMP will be updated.

Table 1: Overview of first cohorts to be incorporated to euCanSHare.

Health initiative	EuCanSHare partner	Number of studies	Aggregated cohort sample size
MORGAM/ BiomarCaRE Project	THL	21	247426
UK Biobank	QMUL	1	500000
SHIP	UMG	1	8728
Hamburg City Study	UKE	1	5000

These cohort studies consist of prospective studies from geographically diverse sources involving a broad range of baseline and follow-up measures and records of clinical outcomes collected from hospitals, clinical and genetic laboratories, health care institutes and administrative record databases or contributed by practitioners or volunteers/ patients themselves through professional-guided or web-based questionnaires**.** For many cohorts, also biological samples are available for future analysis (see Table 2 and Appendix Table 1).

The data and the expertise from these pre-existing initiatives, along with standards and agreements reached at the beginning of the project (WP1), will serve as a reference for data types and formats of interest for the project. These will shape the development of relevant infrastructure and data management model, including the adaptation of repositories to incorporate new data types and formats currently not considered in the centralised data repositories, such as cardiac imaging and general and clinical phenotypic data (See Table 2). This reference will be updated throughout the project to adapt to upcoming state-of-the-art measures, technology and analysis methods.

Data types collected in euCanSHare are expected to include quantitative, qualitative and categorical data generated from surveys, interviews, clinical measurements and tests, medical equipment and sensors. A consensus on open formats to be implemented throughout the project will be reached in WP1.

Table 2 summarises the main categories and data types included in the initial cohorts joining euCanSHare.

For each data type, the relevant variables and required metadata for euCanSHare will be defined on the basis of types/subtypes of these initial cohort data and the expert opinion of cardiovascular researchers and data managers participating in the project (WP3).

Table 2: Overview of data types from initial cohort data incorporating euCanSHare.

Data category	Data type	Variable groups	Sources	Variable types	Transfer formats
General phenotypic data	**socio-demo-graphic**	ethnic background, own and parent´s country/region of birth, country/region of residence, ethnic group, housing conditions, community safety, work settings, income, educational level	questionnaires, interviews, administrative records	numerical, scores, categorical, boolean	csv

General phenotypic data	**lifestyle**	physical activity, sun exposure, smoking and alcohol consumption habits, diet, medication intake, pregnancy, parity, menstruation, contraception, menopause, hormone pills intake	questionnaires, interviews, body sensors	numerical, numerical scores, categorical	csv, cwa
Environmental data	**Environmental**	second-hand smoking pattern and physical environment measures such as residential air pollution, residential noise pollution, greenspace, distance to the coast	questionnaires, interviews, administrative records	numerical, numerical scores, categorical	csv
Clinical phenotypic data	**Medical history**	history of cardiovascular events, diabetes and hypertension and family history of cardiovascular events or heart disease	questionnaires, medical records	numerical, numerical scores, categorical	csv
Clinical phenotypic data	**Physical measures**	blood pressure measures, ECG, height, weight, waist and hip circumferences, cardiopulmonary exercise tests	medical equipment software, medical records	numerical, numerical scores, categorical	csv
Clinical phenotypic data	**clinical assay results**	biochemical parameters such as cholesterol levels, biomarkers, diagnostics of heart disease and diabetes	medical records, clinical lab reports	numerical, numerical scores, categorical	csv
Clinical phenotypic data	**clinical outcomes**	fatal and non-fatal cardiovascular events, diagnosis of heart disease or diabetes, and death	medical records, administrative records	numerical, numerical scores, categorical	csv
Genetic data	**Genotypic/ genomic data**	genotypes, exome-sequencing, genome sequencing and genotypic/genomic-based analysis	sequencing equipment, lab reports	text, numerical, categorical	PLINK, FASTQ, BAM, VCF, CEL, csv
Imaging data	**Cardiovascular-related imaging**	heart MRI, carotid artery US	medical equipment software	image, text, numerical	DICOM, TIFF, csv
Biosamples	**Biosample**	blood (serum, plasma), saliva (mouth swap), DNA	clinical lab	NA	-

*data transfer formats are described in Appendix (section 6.2)

The definition of the minimal required cohort data for a study to enter euCanSHare along with the required metadata for each type of dataset will grow on the basis of this info and in relation to the storage and data management infrastructures and will constitute the submission requirements defined in WP3, which will be stipulated in the submission documentation (See 2.1 Data collection and storage).

2 FAIR Data management

FAIR data management[9] lies at the heart of the euCanSHare project, as euCanSHare will centralise various and heterogeneous cardiology-related data and will integrate different data specialised repositories and analysis tools, maintaining a secure and sustainable platform for responsible data sharing and analysis.

Thus, euCanSHare data management will involve:

- the development of models and protocols for data and metadata submission to specialised repositories
- the adaptation of repositories and platforms for secure storage and reuse of cardiology research data
- the development of an integrated infrastructure to ease data access
- the integration and interoperability of platform elements for storage, access, quality control and harmonisation and data analysis of the heterogeneous cardiology research data

The project´s core will be the implementation of an integrated network of high-quality specialised infrastructures for FAIR research data management that will build on rich metadata vocabularies, ontologies and standards and state-of-the-art data management methodologies for interoperability, and will leverage innovative approaches that will enhance its maximal re-use.

General data management workflow is depicted in Figure 1.

A centralised web portal will constitute the entry point and main interface to the platform, through which the users will have access to documentation of all the main functionalities of the platform (WP2), namely: Coordinated Submission Portal, Cohort Browser, Workspace, Data Access module and Data Analysis module.

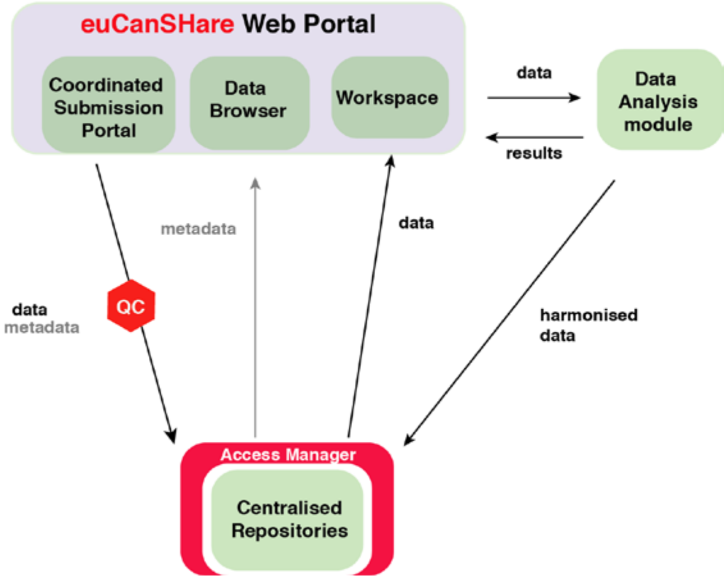

Figure 1: Scheme of data flow within the euCanSHare network. Arrows indicate the direction of data flow. QC denotes central quality control tools.

[9] *H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, 2016.

2.1 Data submission and storage

Data incorporation into euCanSHare will involve the compliance with protocols for data and metadata submission (formats, standards, etc.) defined in WP3 and the centralised deposition of high-volume cardiovascular data of diverse data types in the corresponding repositories.

Data storage is based on the integration of pre-existing specialised repositories, and starts with the coordinated submission of data to such repositories. Some expansion of these platforms could be necessary to accommodate specific cardiology-relevant research data, on the basis of the initial agreements on data types and formats reached by cardiology and data management experts (WP2-4).

Cohort data entering euCanSHare will be distributed according to their nature to be physically stored in specialised repositories. Data types and formats that will be stored in each repository along with required metadata is described in Table 3.

Table 3: Overview of repositories of data and metadata in euCanSHare.

General data type	Central repository	Data formats
Metadata	EGA	formats to be defined in WP3
Genotypic and omic data	EGA	encrypted files; all standard formats of sequence and array-based omic data and derived analysis data
Phenotypic (socio-demographic, lifestyle, etc)	EGA	encrypted files; formats to be defined in WP3
Cardiac imaging	euro-BioImaging	anonymised files; Dicom, TIFF, csv for image-derived data such as segmentation and meshes, and other associated quantitative data (under development)
Biosamples	BBMRI	serum, plasma, saliva, DNA, RNA

The Coordinated Submission Portal will centralise documentation and guidelines covering the minimal conditions of cohort data and required metadata for a cohort study to join euCanSHare, the description of the procedure for evaluation/approval of study´s requests to join euCanSHare and the submission requirements and procedures for data deposition in the central specialised repositories (WP2).

As part of the submission procedure, submitters are required to define an access policy for each dataset.

Unique permanent identifiers will be generated for each study and specific dataset, which will allow cross-linkage within the network.

Metadata will be stored in the European Genome-phenome Archive (EGA), a core data resource and deposition database for biomolecular data from ELIXIR. Metadata is submitted through the EGA Submitter Portal https://ega-archive.org/submission/) or programmatically via a REST API (using JSON or XML formats).

Phenotypic and genetic/omic data will be stored in EGA and should accomplish the requirements specified in EGA submission documentation, including compliance with the informed consents and the anonymization sensitive data previous to submission. The submission request is evaluated, and after approval, submitters will be guided to comply with the specifications of data submission, including the download, installation and use of software for encryption of data. Data submitted to the EGA must be submitted along with data-type specific required metadata (Table 3). Encrypted data (phenotypic and genetic data) should be transferred from the submitter premises to the EGA through the available file transfer technologies (e.g. FTP or Aspera). DACs are also created during the submission process, so the resulting datasets are associated to an access committee. Unique permanent identifiers are generated for each study (EGAS), specific dataset (EGAD) and DAC (EGAC), actually for every relevant object.

Imaging and **imaging-derived** data will be submitted to Euro-BioImaging (http://www.eurobioimaging.eu/), the ELIXIR's partner for the Image Data Strategy, via UI or APIs on Euro-BioImaging platform and should accomplish the requirements specified in the submission documentation, including the anonymization of data previous to submission. Data will be sent via an encrypted channel to the XNAT archive-based central repository. Unique identifiers will be assigned to data samples and datasets.

Bio-samples will be deposited in geographically distributed dedicated biobanks from the Biobanking and BioMolecular resources Research Infrastructure (BBMRI)[10] according to the guidelines and quality management/assurance protocols following international CEN Technical Specifications and ISO 20387:2018 international standards on sample generation, processing, anonymisation and transporting procedures. Upon submission, persistent unique identifiers are assigned to samples and subjects/donors and associated required and additional metadata are entered to the BBMRI catalogue.

2.2 Data flow within the network

The components of the network will include:

- a Data Browser module (Cohort Browser) that will allow users to locate data of their interest
- a Data Access subportal to request access to selected data
- an Access Manager that will handle actual access to the data
- a Workspace that will allow users to submit data they have been granted access to for quality control, harmonisation or analysis to visualize and download results
- a Data Analysis module

A Data Manager will integrate the networked components for data storage and analysis (see Table 4).

The Data Analysis module will be based on previous infrastructure developments, established industry standards (cloud managers, Docker container technology, Galaxy, Common Workflow Language), and accepted protocols and will follow the recommendations of Global Alliance for Genomics and Health (GA4GH)[11], ELIXIR[10], and will assure compatibility in protocols and standards with the European Open Science Cloud (EOSC) initiative[12].

Data and metadata storing platforms (EGA, Euro-BioImaging and BBMRI) will implement mechanisms for linking euCanSHare study´s data and metadata across platforms based on the unique and persistent study, datasets and sample identifiers assigned upon submission (See 2.1 Data submission and storage). These identifiers will allow linking the studies that have been accepted to join euCanSHare into the euCanSHare Cohort Browser and to the corresponding analysis tools/ platforms from the Data Analysis module (See figure 2 and table 4).

[10] Litton J-E. Launch of an Infrastructure for Health Research: BBMRI-ERIC. Biopreserv Biobank 2018;16(3):233–241.

[11] Terry SF. The Global Alliance for Genomics & Health. *Genet Test Mol Biomarkers* 2014;18(6):375–376.

[12] Mons B, et al., Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 2017;37(1):49–56.

Metadata will be available from the **Cohort Browser**, while data will flow through more strict access control rules (See 2.5 Data Accessibility) to the private workplace of authorised researchers and through the same authorization to the analysis tools.

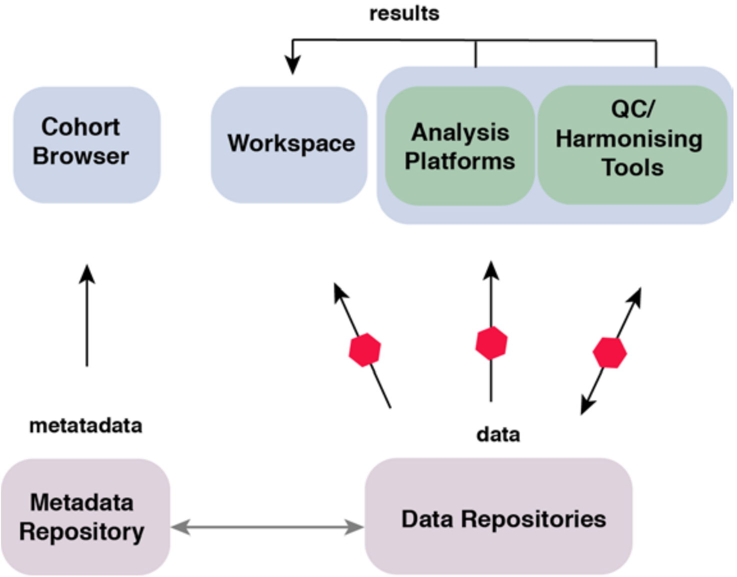

Figure 2: Flow of data and metadata within the network. Metadata is readily findable from the Cohort Browser as a means to discover cohort data of interest. Data flow from data repositories to the private workspace of authorised users after DAC approval and through their credentials to components of the Data Analysis module are indicted through the arrows. Hexagons represent controlled data flow.

Once access is granted , data and access credentials will be linked to user´s private cloud-based **Workspace** from which data will flow under specific access control rules from the relevant repositories to their corresponding quality control/harmonising/analysis tools within the Data Analysis module (Table 4)

Table 4: Flow of Metadata and data between repositories and their corresponding QC, harmonisation tools and analysis tools/platforms within the network for each data type.

Data type	Centralised repository	Quality Control / Harmonisation tools	Analysis tools/platforms
Metadata	EGA	Mica	-
Genetic data	EGA	ELIXIR Bioinformatics toolbox (T4.4)	ELIXIR Bioinformatics toolbox (T4.4)
Phenotypic (socio-demographic, lifestyle, etc)	EGA	Square2 / Opal	all
Cardiac imaging	Euro-BioImaging	new tools from WP4	Radiomics toolbox (T4.2)
Bio-samples	BBMRI	BBMRI services	-

EuCanSHare will set up a **Data Access module** (comprising a Data Access subportal and a Data Manager infrastructure), dedicated to facilitating the process of acquiring access credentials according to the access rules of the respective cohorts/ datasets, defined and stored upon data submission (See 2.1 Data collection and storage).

The Data Access subportal will host and disseminate access policies and procedures aligned to GA4GH Standards (using Data Use Ontologies (DUOs), derived from Data Use Condition (DUC) and Automatable Discovery and Access Matrix (ADA-M). The Data Access subportal will provide a simple framework for researchers to apply for data access to different data.

The Access Manager component will manage the procedures for granting access, will handle the data requests from the Data Access subportal and will manage granted credentials thereby allowing data to flow within the network in a controlled manner. To accomplish this, the Access Manager will leverage on strategies already being implemented in the EGA infrastructure, namely the Data Access Committees (DACs: https://ega-archive.org/dacs) to centralise and facilitate the procedure of granting access and managing granted credentials. The possibility of automatic credentials assignment based on applications and policies metadata will be explored through a blockchain technology as will be the developing of a tool similar to the Data Use Oversight System (DUOS) from Broad Institute (https://duos.broadinstitute.org/#/home) that could be implemented as a system to pre-process applications before submitting them to the Data Access Committees (DAC) to improve/facilitate the access procedure.

The **Data Analysis module** will be implemented as a virtual environment to perform analysis on the data under the appropriate security conditions (WP4).

This module will integrate infrastructure for data harmonisation and data quality control solutions, and a comprehensive toolbox for bioinformatics and imaging analysis for cardiovascular personalised medicine research (WP4)(Table 3), with interoperable customisable pipelines within the framework of the well-established environment Galaxy[13] and the Docker container technology (https://www.docker.com/) for bioinformatics and the FASTR platform for imaging analysis[14].

As data flows through authorised user's credentials from the relevant repositories to and between the appropriate platforms (Table 3) researchers will be able to build computational workflows on the selected cohorts including and combining image, phenotypic and omic machine learning/statistical methods to quantify associations and new biomarkers, and to build predictive models of cardiovascular disease.

Harmonisation algorithms used for cardiovascular cohorts and variables will be stored in a standardised electronic database within a centralised web-based harmonisation management system allowing study coordinators and data managers to securely import/export a variety of cardiovascular data types and harmonisation rules in different formats.

Upon request, harmonised datasets and analysis results will be transferred into the relevant data repositories associated with new persistent and unique identifiers as a related submission with the appropriate links to original and related versions, allowing its readily discoverability and re-usability. Quality control and analysis results and will be transferred to the workspace where they will be available for authorised researchers.

2.3 Interoperability

A consensus on data types and formats will be adopted to shape the communication channels among euCanSHare data components and to implement a common data model seeking the optimal interoperability at the syntactic level.

[13] Afgan E, et al., The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 2018;46:W537–W544.

[14] Achterberg HC A, Koek M and Niessen WJ. Fastr: A Workflow Engine for Advanced Data Flows in Medical Image Analysis. Front ICT 2016;3(15).

Metadata standards for specific data types commonly used in cardiovascular research will be adopted for imaging, omics, epidemiological, clinical data, and bio-samples and the harmonisation models implemented during the MORGAM/BiomarCaRE project and emerging solutions (Opal/Mica) will be leveraged on for the harmonisation of standard vocabularies, naming conventions and ontologies through the building of a standardised electronic database for harmonisation algorithms for cardiovascular cohorts and variables.

Harmonisation, quality control and analysis platforms tools will be integrated within customised pipelines built through the well-established work flowing environment Galaxy[15] (www.usegalaxy.eu) and the Docker container technology (www.docker.com) for bioinformatics and FASTR platform[16] for imaging analysis.

2.4 Making data findable

Data in EuCanSHare project will be findable through a Data Browser in the Web Portal (See section 2.2).

Efforts of cardiology research experts, data contributors, data managers and analysis platform´s experts will lead the implementation of metadata standards for the specific needs of the project (data types, sources, technologies) along with the definition of required metadata for the different data types, variable naming conventions and data types categories. This will allow the assembly a comprehensive cardiovascular data catalogue, with a rich set of metadata and keyword dictionaries for data types and categories accounting for state-of-the-art technology, methodologies and data analysis procedures in cardiology research (WP2-3).

This data catalogue will allow the euCanSHare Cohort Browser to implement a powerful searching engine allowing researchers to quickly find the information they need to identify studies or datasets of interest and locate variables and data they need for implementing cardiovascular research projects by means of descriptive metadata and widely used classification formats (e.g. ICD-10-CM codes for classification of diseases and diagnoses, SOC2003 coding for occupations).

Persistent and unique identifiers assigned upon submission to metadata and data repositories will be linked together and to the Cohort Browser in order to ease the location of studies, datasets and samples and discoverability of studies in an integrated manner. Also, versions of datasets and related datasets including harmonised datasets and analysis results could be stored into data repository associated with new persistent and unique identifiers and the appropriate links to original and related versions, allowing its readily discoverability and re-usability.

2.5 Data accessibility

EuCanSHare Web Portal will provide public metadata both by interactive and programmatic access (using REST protocol for data access and Oauth2 for authentication) as well as the necessary information about the protocol to access protected data.

Access to data will be provided to researchers in compliance with access policies that observe the informed consents from participants, through the Data Access module (see 2.2.3 Access Manager). After request approval by the corresponding DAC, the management of granted credentials by the access management system will be allow authorised researchers to access data from their Workspace and data to flow through user-authenticated and secure channels within the network (See 2.2 Data flow within the network).

All software needed for access to data and analysis will be embedded in computational infrastructure along with appropriate help and tutorials.

2.6 Increase data re-use

EuCanSHare will provide an infrastructure to maximize the responsible re-use of cardiology research data within the ethical and legal framework for compliant responsible data sharing (WP1), including:

- A comprehensive catalogue of relevant metadata in line with the standards of cardiology research

- A central storing platform made of certified data-type specialised repositories will allow the secure long-term storage and re-use of high-quality data.

- A central web portal, allowing secure and user-friendly access to the platform functionalities.

- A comprehensive Cohort Browser within the Web Portal that will ease the discoverability of cardiology data.

- Integrated data repositories and state-of-the-art quality control harmonising and analysis software based on powerful high-performance computing and compute cloud.

- Easily accessible documentation, training and user support resources (e.g. guidelines, tutorials and workshops informing users on methods, resources and policy tools offered) will encourage the use the functionalities of the platform, ensuring the reuse of the platform for accessing information and perform analysis and to increase the network by depositing new data and tools to persist well beyond the duration of individual projects.

- Infrastructure to easily incorporate new features such as visualization and analysis tools into the platform will allow keeping the platform at pace with evolving technology and analysis methods.

- A central data quality control assurance approach leveraging on metadata handling and automatised data monitoring processes (WP4 T4.2).

- The storage of harmonisation algorithms in a standardised electronic database such that any harmonisation effort can be easily searched and located in the database from the Cohort Browser and re-used in new multi-cohort research studies.

3 Data security

Metadata and data will be stored in specialised repositories (EGA and Euro-BioImaging) that are duplicated and backed-up.

Transfer of sensitive data within the network will be assured through encryption protocols, a secure communication interface and encrypted channels for securely connecting the data to the relevant euCanSHare users and to data processing environments including cloud-based execution tools. This will involve easy-to-use modules for authentication (based on Oauth2) and secure data management.

Biosamples will be stored in institutionalised biobanks following international standards for quality assurance including provisions for safe and secure long-term preservation of human samples.

4 Allocation of resources

Metadata repositories in euCanSHare will build upon well-developed European-funded specialised platforms of distributive nature (ELIXIR/EGA, EuroBioImaging, BBMRI) which will assure long-term safe storage and sharing of data. Extension of these repositories to improve FAIR cardiology research data and metadata sharing in line with state-of-the-art analysis methodologies and tools will be accomplished as part of this project.

5 Appendix

5.1 Description of Initial cohort studies

Table 5: Overview of initial studies joining euCanSHare.

Study	Partner	Data owner	Origin	Source	Collection methodology	Cohort composition	Size	Data types
Alpha-Tocopherol Beta- Carotene Prevention (ATBC)	THL	Department Health, National Institute for Health and Welfare, Helsinki, Finland	Southern and Western Finland (Nationwide Central Population Register)	Register of Causes of Death; Hospital Discharge Register; Individual study participant	Baseline self-reported/documented history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events.	1 cohort of men aged 49-70 years at recruitment, derived from the Nationwide Central Population Register	29133	lifestyle data; physical measurements; sociodemographic; clinical outcomes; genotypic data; biosamples; environmental;
Caerphilly Prospective Study	THL	Population Health Sciences, University of Bristol, Bristol	Caerphilly. It includes Caerphilly, a former mining town in South Wales, and outlying villages	National Health Service (NHS) Registry and Office of National Statistics (ONS); Hospital Episode Statistics (HES); Hospital notes; General practitioners' records; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events.	1 cohort of men aged 52-72 years at recruitment	2171	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples
The ESTHER Study	THL	German Cancer Research Centre, Division of Clinical Epidemiology and Aging Research, Heidelberg, Germany	Saarland. The federal state Saarland in South-West Germany	Population Registers; Regional Health Departments; General practitioner; Saarland Cancer Registry; individual participant	Baseline self-reported history of cardiovascular events, documented history of diabetes and cancer and follow up of fatal and non-fatal cardiovascular events, diabetes and cancer.	1 cohort of men and women aged 48-75 years at recruitment	9949	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples
Estonian Genome Centre University of Tartu	THL	MORGAM Participating Centre 81: Estonia	Estonia. It includes the whole country.	Estonian Causes of Death Registry; Databases of the Estonian regional and central hospitals; The Estonian Cancer Registry; Estonian Health Insurance Fund; Estonian Myocardial Infarction Registry; Population Registry; Individual study participant	Baseline self-reported/documented history of cardiovascular events and documented history of diabetes and cancer and follow up of fatal and non-fatal cardiovascular events.	1 cohort of men and women aged 18-84 years at recruitment	4971	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; environmental

The National FINRISK Study	THL	Centre for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, The Capital Region of Denmark, Copenhagen	Former province of North Karelia; Former province of Kuopio; Town of Helsinki, the capital of Finland and the adjacent town of Vantaa; Town of Turku and Loimaa district; Oulu Province; Lapland Province.	Register of Causes of Death; Hospital Discharge Register; FINMONICA coronary event and stroke registers; INAMI and FINSTROKE register; FINMONICA-FINAMI register of cardiac revascularizations; Drug reimbursement registers; Individual study participant	Baseline self-reported/ documented history of cardiovascular events, documented history of cancer and diabetes and follow up of fatal and non-fatal cardiovascular events, cancer and diabetes	5 cohorts of men and women aged 24-74 years at recruitment from different reporting units and timespans	38333	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; genotypic data; metabolites; environmental
DAN-MONICA	THL	Helmholtz Zentrum München - German Research Centre for Environmental Health, Neuherberg, Germany	Eleven municipalities in the western part of Greater Copenhagen Area.	Civil Registration System (CRS); Causes of Death Register; National Hospital Discharge Register; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events and diabetes	3 cohorts of men and women aged 29-71 years at recruitment from different reporting units and timespans	7582	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; metabolites; environmental
Kooperative Health Research in Region Augsburg (KORA)	THL	Department of Health Promotion and Chronic Disease Prevention, National Public Health Institute, Helsinki, Finland	Augsburg. City of Augsburg and the less urban Landkreis Augsburg and Landkreis Aichach-Friedberg in Bavaria, Southern Germany.	Population Registries; Regional Health Departments; Coronary Events Registry; Hospitals and general practitioners; Individual study participant	Baseline self-reported history of cardiovascular events and diabetes and follow up of fatal and non-fatal cardiovascular events and diabetes	4 cohorts of men and women 24-75 years old at recruitment from different timespans	17264	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; genotypic data; metabolites
Moli-Sani Project	THL	Department of Epidemiology and Prevention, IRCCS Mediterranean Neurological Institute – NEUROMED, Pozzilli (IS)	Molise. Molise region in central Italy.	Municipalities; Death Registry of the National Health Service (ReNCaM); The National Health Service; Hospital Discharge Records; General practitioners; Individual study participant	Baseline self-reported history of cardiovascular events and diabetes and follow up of fatal and non-fatal cardiovascular events and diabetes	1 cohort of men and women aged 35-99 years at recruitment	24325	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; metabolites;environmental
MONICA Brianza	THL	Centro Ricerche EPIMED - Epidemiologia e Medicina Preventiva, Dipartimento di Medicina Clinica e Sperimentale. Università degli Studi dell'Insubria, Varese	Brianza. Seven health districts in historical area between Milan and Swiss border.	Municipality; Hospital discharge records; Individual study participant	Baseline self-reported history of cardiovascular events and diabetes and follow up of fatal and non-fatal cardiovascular events and diabetes	3 cohorts of men and women aged 25-65 years at recruitment from different timespans	4932	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; genotypic data;

								metabolites; environmental
MONICA Catalonia	THL	Institute of Health Studies, Department of Health, Barcelona	Catalonia. Central area of Catalonia (Counties of Bages, Berguedà, Solsonés, Vallés Oriental and Vallés Occidental, which include a part of the Barcelona metropolitan area and up to the Pyrenées)	Regional mortality register; National Death Index; Coronary Event Register; Municipal population register; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events	2 cohorts of men and women aged 24-67 years at recruitment from different timespans	5505	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples
MONICA Friuli	THL	Centro Ricerche EPIMED - Epidemiologia e Medicina Preventiva, Dipartimento di Medicina Sperimentale. Università degli Studi dell'Insubria, Varese; Cardiovascular Prevention Centre, ASS4 Medio Friuli and Agenzia Regionale della Sanità, Udine	Three provinces of the Friuli-Venezia Giulia region of north-east Italy, bordering Austria and Slovenia	Regional Health Information System; General Registry Office; Individual study participant	Baseline self-reported history of cardiovascular events and diabetes and follow up of fatal and non-fatal cardiovascular events	3 cohorts of men and women aged 24-64 years at recruitment from different reporting units and timespans	5510	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; environmental
MONICA Kaunas	THL	Institute of Cardiology, Kaunas University of Medicine, Kaunas, Lithuania	Kaunas city	Population-Based Mortality Register; Coronary Event Register; Stroke Event Register; Kaunas Address Bureau; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events	3 cohorts of men and women aged 33-65 years at recruitment from different timespans	4485	lifestyle data; physical measurements; sociodemographic; clinical outcomes; environmental
MONICA Newcastle	THL	Faculty of Health, University of Newcastle, New South Wales, Australia	Newcastle; Lake Macquarie; Maitland; Cessnock; Port Stephens	National death index; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal events	3 cohorts of men and women aged 34-70 years at recruitment from different reporting units and timespans	5873	lifestyle data; physical measurements; sociodemographic; environmental

MONICA Northern Sweden	THL	MORGAM Department of Medicine, University Hospital, Umeå, Sweden	Västerbotten County; Norrbotten County	National Death Register; National registers at the National Board of Health and Welfare; Local Population Registers; Local Diagnosis Registers; Myocardial Infarction (MI) and Stroke Event Registers; Diabetes register; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events, diabetes and cancer	7 cohorts of men and women aged 24-75 years at recruitment from different reporting units and timespans	12013	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; environmental
MONICA Novosibirsk	THL	MORGAM Participating Centre 47: Institute of Internal Medicine	Octyabrsky District and Kirowsky District of Novosibirsk city	Population-Based Mortality Register of the Institute of Internal Medicine; Coronary and Stroke Event Registers; Individual study participant	Baseline self-reported history of cardiovascular events and follow up of fatal and non-fatal cardiovascular events	3 cohorts of men and women aged 24-65 years at recruitment and 1 cohort of men 23-63 years at recruitment from different reporting units and timespans	11438	lifestyle data; physical measurements; sociodemographic; clinical outcomes; environmental
MONICA Tarnobrzeg	THL	MORGAM Participating Centre 35 - Krakow: Department of Epidemiology and Population Studies, Institute of Public Health, Faculty of Health Care, Medical College Jagiellonian University, Krakow	Tarnobrzeg Voivodship.	Voivodship Death Register; Local Address Register; individual study participant or their relatives	Baseline self-reported history of cardiovascular events and follow up of fatal cardiovascular events.	3 cohorts of men and women aged 34-65 years at recruitment from different timespans	5362	lifestyle data; physical measurements; sociodemographic; environmental
MONICA Warsaw	THL	MORGAM Participating Centre 36: The Cardinal Stefan Wyszynski Institute of Cardiology, Warsaw, Poland	Warsaw Praga South, Warsaw Praga North (districts of Warsaw)	Polish Universal Electronic Population Register (PESEL); Central Death Register; MONICA Coronary and Stroke Registers; individual study participant	Baseline self-reported history of cardiovascular events and diabetes and follow up of fatal and non-fatal cardiovascular events	3 cohorts of men and women aged 34-65 years at recruitment from different timespans	5577	lifestyle data; physical measurements; sociodemographic; clinical outcomes; environmental
Prospective Study of Myocardial Infarction (The PRIME study)	THL	PRIME/Belfast (Queen's University of Belfast, Centre for Public Health, Institute of Clinical Science, Northern Ireland, UK); PRIME/Lille (INSERM U 744-Institut Pasteur de Lille); PRIME/Strasbourg (Department of	Belfast; Toulouse; Lille; Strasbourg	General practitioner, specialist or occupational medicine department; Registrar General's data; Business Services Organisation (BSO); Northern Ireland Cancer Registry; individual study participant or their relatives	Baseline self -reported history of cardiovascular events, cancer and diabetes and follow up of fatal and non-fatal cardiovascular events, cancer and diabetes	1 cohort from Belfast of men aged 49-60 years old men; 1 cohort from Lille of men aged 50-59 years; 1 cohort from Toulouse of men aged 49-60 years	10600	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples; genotypic

		Epidemiology and Public Health, University of Strasbourg); PRIME/Toulouse (Department of Epidemiology, National Institute of Health and Medical Research, Unit 558)				old men; 1 cohort from Strasbourg of men aged 50-59 years.		
The Tromsø Study	THL	Institute of Community Medicine, University of Tromsø	Tromsø	Local Population Registers; Central Population Register (DSF); Causes of Death Registry; The University Hospital of North Norway	Baseline self -reported/ documented history of cardiovascular events, cancer and diabetes and follow up of fatal and non-fatal cardiovascular events	1 cohort of men and women aged 19-62	31847	lifestyle data; physical measurements; soci odemographic; clinical outcome; biosample; environmental
MATISS (Malattie Ateroscleroticheh e Istituto 31 Superiore di Sanità) study	THL	Cardio-Cerebrovascular Epidemiology Unit, National Centre of Epidemiology, Surveillance and Health Promotion, Istituto Superiore di Sanità, Rome, Italy	Four municipalities (Priverno, Sezze, Bassiano and Roccagorga) of the Region of Latina.	Municipalities; Death Registry of the National Health Service; Coronary and Cerebrovascular Event Register; Hospital Discharge Records; Person himself/Re-examination; general practitioner; individual study participant	Baseline self-reported questionnaires and follow up of fatal and non-fatal cardiovascular events.	3 cohorts of men and women aged 18-81 years at recruitment from different timespans	8512	lifestyle data; physical measurements; sociodemographic; clinical outcomes; biosamples
PAMELA Arterial Pressure Study	THL	Brianza Centro Ricerche EPIMED - Epidemiologia e Medicina Preventiva, Dipartimento di Medicina Sperimentale. Università degli Studi dell'Insubria, Varese; Dipartimento di Medicina, Prevenzione e Biotecnologie Sanitarie. Università degli Studi Milano-Bicocca, Monza	City of Monza.	Municipality; Hospital discharge records	Baseline self-reported questionnaires and follow up of fatal and non-fatal cardiovascular events.	1 cohort of men and women aged 25-75 years at recruitment from different reporting units and timespans	2044	lifestyle data; physical measurements; sociodemographic; clinical outcomes
Study of Health in Pomerania (SHIP and SHIP-TREND)	UMG	Research Network Community Medicine (FVCM)	West Pomerania, a region in the northeast of Germany.	SHIP associated medical centres and clinical labs	Trained and certified interviewers and computer-assisted personal interviews, audited regularly; clinical laboratory data obtained according to standardised procedures; genotyping using SNP 6.0 Arrays;	2 cohorts of men and women aged 20– 79 years from different timespans	8728	sociodemographic; biosamples; genotypic data; cardiac imaging; lifestyle; environmental; clinical outcome sociodemographic;

								biosamples; genotypic data; cardiac imaging; lifestyle; environmental clinical outcome sociodemographic; biosamples; genotypic data; cardiac imaging; lifestyle; environmental clinical outcome
UK Biobank	QMUL	Board of UK Biobank	UK	United Kingdom's National Health Service (NHS) centres	Following of participants by surveys, accessing their health records and national registries and monitoring physical activity.	Middle-aged population	500000	socio-demographic, family history, lifestyle, medical history, biological measurements, lifestyle indicators, biomarkers and imaging
Hamburg City Health Study	UKE	University Medical Centre Hamburg	Hamburg City	Population based study, invited to participate	Following of participants by examination at study centres, questionnaires	1 cohort of inhabitants from Hamburg, aged 45-74 years at recruitment	5000	socio-demographic; environmental; physical measures; lifestyle; biosamples; omics and imaging

5.2 Definition of data transfer files for initial cohort studies

BAM: Binary Sequence Alignment/Map

CEL: data file created by Affymetrix DNA microarray image analysis software

csv: comma-separated values format

cwa: Continuous Wave Accelerometer format, binary packed format for raw acceleration data from wearable devices

DICOM: Digital Imaging and Communications in Medicine, the international standard to transmit, store, retrieve, print, process, and display medical imaging information.

FASTQ: ASCII text-based format for storing both a biological sequence and its corresponding quality scores

PLINK: files for Genome Wide Association Studies

TIFF: Tag Image File Format

VCF: Variant Call File