euCanSHare Deliverable D3.1:
Data Management Plan v1

EGA-CRG

April 5, 2019

Abstract

EuCanSHare aims at harboring the most comprehensive cardiovascular data catalogue ever assembled by integrating major cardiovascular data sources from Europe and Canada, and facilitating enhanced cross-border data sharing, discoverability and exploitation for personalized medicine research in cardiology. Thus, FAIR data management is at the heart of euCanSHare project, as euCanSHare will store/centralise data coming form a variety of cardiovascular studies, maintaining a secure and sustainable platform for sharing and analysis.

This Data Management Plan addresses the purpose and description of data handled within the euCanSHare project and the implementation of a model for data handling during and after the project, from data deposition/collection, preservation and curation, data integration and interoperability, accessibility and exploitation, including standards and methodology applied.

Contents

1 Data Summary

1.1 Purpose of data collection

Project proposal

- State the purpose of the data collection/generation
- Explain the relation to the objectives of the project
- Outline the data utility: to whom will it be useful

Q>
1-What is the purpose of the data collection/generation and its relation to the objectives of the project?
2-Will you re-use any existing data and how?
2-To whom might it be useful ('data utility')?

The data collection is central to the general purpose of euCanSHare project, which aims at centralizing and securing cardiology medical/research data while providing a sustainable platform for cross-border data sharing and multi-cohort analysis.

Datasets included in euCanSHare are meant for re-usability and re-purpose through a Web Portal and Interoperability and Analysis Interfaces and so are planned to become a central tool for cardiology personalized medicine. How will data will be re-used?. Outline of protocols here?

As personalized medicine approaches are urgently needed in cardiovascular research to improve risk assessment and early diagnosis, as well as for treatment personalization and drug development, data collected in euCanSHare project along with their analysis available from euCanSHare's Analysis Platform have potential interest for both researcher and medical staff.

Data will be reusable through a centralised web portal to the platform, which, in addition to the data repositories or their links the platform will integrate a range of established technologies developed by our consortium members (cf. Table 2 of project proposal) to offer comprehensive computing infrastructures to its users, encompassing software (for cataloguing, harmonisation, co-analysis and storage of study-specific and harmonised data); methods (supporting complex data integration models); high-performance computing and compute cloud (for data processing and storage); training and user support resources (e.g. guidelines, tutorials and workshops informing users on methods, resources and policy tools offered), and a central web portal (allowing secure and user-friendly access to the platform catalogues and functionalities). Figure 2 of project proposal represents the main components and functionalities of the platform.

1.2 Description of data

Project proposal/Datasets contributors/ experts> Tarja Palosaari and Ari Haukij□rvi

- Specify the origin of data generated/collected
- Specify the types and formats of data generated/collected
- State the expected size of the data (if known)

Q>
1-What is the origin of the data?
2-What types and formats of data will the project generate/collect?
3-What is the expected size of the data?

EuCanSHare will build on existing databases from established multi-cohort initiatives in cardiology (MORGAM, BiomarCaRE, CAHHM) as well emerging comprehensive databases such as the UK Biobank and the Hamburg City

Health Study (see Table 1 of project proposal) and newly submitted datasets from contributors from Europe and Canada.

In an initial phase euCanSHare an initial reasonably diverse set of cohorts (35 cohorts) will be integrated into the catalogue and data repositories, based on four different sources, namely:

1. The MORGAM data (www.thl.fi/publications/morgam/cohorts/full/contents.htm), managed by THL;
2. BiomarCaRE cohorts (www.biomarcare.eu), coordinated by UKE;
3. The Canadian Alliance CAHHM (www.cahhm.mcmaster.ca) coordinated by MCM;
4. Other major cohorts not yet integrated into these networks such the UK Biobank (QMUL) or the highly rich Study of Health in Pomerania ? SHIP (UMG).

Data types included in datasets include socio-demographics, bio-samples, omics, cardiac images, lifestyle data, environmental data, physical measures and clinical outcomes. It is expected that new data types may be submitted? Ask datasets contributors, experts?
Instead of a general enumeration of data types and its posible formats included in cohorts a table showing all data types and formats per study- ask contributors for info

See Standsrds for data type naming. Also variables per data type?

Table 1: Data types and its formats
(from initial set of cohorts)

Data type	Formats
socio-demographics	surveys?, many things, more things
bio-samples	things, many things, more things
omics	things, many things, more things
cardiac images	things, many things, more things
lifestyle data	things, many things, more things
environmental data	things, many things, more things
physical measures	things, many things, more things
clinical outcomes	things, many things, more things

1.3 Data security levels, confidentiality of potentially disclosive personal information

Ask each contributor

- Specify different levels of data security>open (aggregated statistics,demo?), controlled

Table 2: Data types and its needs for long-term storage, security issues, require metadata and interoperability requirements

(from initial set of cohorts)

Data type	Storage	Security	Metadata	Interoperability
socio-demographics	long-term needs	security issues	required metadata	required io method
bio-samples	long-term needs	security issues	required metadata	required io method
omics	long-term needs	security issues	required metadata	required io method
cardiac images	long-term needs	security issues	required metadata	required io method
lifestyle data	long-term needs	security issues	required metadata	required io method
environmental data	long-term needs	security issues	required metadata	required io method
physical measures	long-term needs	security issues	required metadata	required io method
clinical outcomes	long-term needs	security issues	required metadata	required io method

2 Data management, documentation and curation

Jordi Rambla, Aad van del Lugt, Marcel Koek

Data management is depicted in the data workflow diagram Data Workflow figure 7 Deliverable D7.1.

Data management will deal with integrating established infrastructures (ELIXIR, EGA, BBMRI, euro-BioImaging) for managing storage of the heterogenous cardiac data, tools/procedures for uploading new cohorts into the platform, and (iii) APIs for securely connecting the data to the relevant euCanSHare users and/or data processing environments (e.g. ELIXIR). Metadata across the different repositories will be linked (-omics, clinical, imaging and bio-samples) to assure the transversal consistency of the specifications, thus presenting the heterogeneous cardiac data to the users in an integrated manner.

The Data Manager will build on euCanSHare?s Data Management Plan (DMP), which will include a comprehensive analysis of the nature of data to be handled, the requirements for interoperability, long-term storage and security issues, as well as compliance with FAIR and EOSC data principles

2.1 Data collection / generation

Jordi Rambla

- Specify the methodology of data deposition for user
- Define protocols for depositing new cohort raw data into the appropriate repositories and methods to provide rich metadata to foster a quality re-use of raw data for newly coming research projects.

Data collection,..deposition method

2.2 Data storage

Jordi Rambla, Aad van del Lugt, Marcel Koek

Data storage will be carried mainly through the a series of data and metadata repositories of distributive nature, including EGA (omics), euro-BioImaging (cardiac imaging), MORGAM/BiomarCaRE (clinical/lifestyle data) and BBMRI (biosamples) and data transfer methodologies (REF)

Storage

EGA> main distributor? The platform will leverage on the European Genome-phenome Archive (EGA)'s infrastructure (figure 8 Deliverable D7.1), a service specialized in permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects, and incorporating some advanced components for phenotype information (similar to those of the American counterpart dbGaP).

Genomic data

Phenotypic data

Image Data will be handled by euro-BioImaging? Image data will be in a central repository based on XNAT?? For those contributors/cohorts not willing to have a central imaging archive, links to other repositories will be provided.

?To guarantee anonymization of personal data, dicom? headers will be removed. ECM? will assist in tthis. A link to the analysis tool will be also provided.

clinical/lifestyle data

- Specify the methodology of data storage
- Data workflow diagram

2.3 Data sharing and access

- Specify if existing data is being re-used (if any)
- Specify technology for data access and transfer
- Specify methodology of data re-use (analysis)

Researcher will be able to work on a workspace from their systems without data download or software installation?? Software on platform?

(iii) APIs for securely connecting the data to the relevant euCanSHare users and to data processing environments (e.g. ELIXIR).

Additionally, EGA is collaborating with the BROAD Institute, Harvard to develop a tool similar to their Data Use Oversight System (DUOS) that is planned to be implemented a as system to pre/process applications before submitting them to the Data Access Comittees (DAC), to improve/facilitate? current access procedure ask Jordi. New functionalities for data management in EGA are to be developed within the Excelerate project, specially for dealing with phenotypic data? Excelerate project figure 9 Deliverable D7.1

See Data security for methodology of secure data storage and transfer of sensitive data.

3 FAIR data management

3.1 Making data findable, including provisions for metadata

metadata- Carsten Oliver Schmidt

- Outline the discoverability of data (metadata provision)
- Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?
- Outline naming conventions used
- Outline the approach towards search keyword
- Outline the approach for clear versioning

- Specify standards for metadata creation (if any). If there are no standards in your discipline describe what type of metadata will be created and how
- Project-Describe relevant models, tools and infrastructures for (meta) data sharing and distribution within the network
- Project-Describe relevant models, tools and infrastructures for (meta) data sharing and distribution within the network

Q>
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)?
What naming conventions do you follow?
Will search keywords be provided that optimize possibilities for re-use?
Do you provide clear version numbers?
What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

3.2 Making data openly accessible

API: Marcel Koek, DAC: Aad van der Lugt

- Specify which data will be made openly available? If some data is kept closed provide rationale for doing so
- Specify how the data will be made available
- Specify 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)?
- Specify where the data and associated metadata, documentation and code are deposited
- Specify how access will be provided in case there are any restrictions

Q>
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.
2-How will the data be made accessible (e.g. by deposition in a repository)? What methods or software tools are needed to access the data?
3-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)?
4-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.
5-Have you explored appropriate arrangements with the identified repository?
6-If there are restrictions on use, how will access be provided?
7-Is there a need for a data access committee?
8-Are there well described conditions for access (i.e. a machine readable license)? How will the identity of the person accessing the data be ascertained?
Three access levels> public data, registered data (need authentication), controlled data (need access permission).

Rules should apply also to users registered rights (documentation) Mahsa Shabani
Data access API wiil be via many endpoints Marcel Koek

DAC will grant permission specific to every cohort Aad van der Lugt

Access policies (in close collaboration with WP1) will be defined and stored in the Automatable Discovery and Access Matrix (ADA-M) introduced by the GA4GH.

acces ADA-M? to ease the process of requesting access to the different cohorts and acquiring access credentials. The task will leverage strategies already being assayed in the EGA infrastructure, namely the Data Access Committees (DACs: www.ebi.ac.uk/ega/dacs).Through MCGILL, one of the original drivers of the Global Alliance for Genomics and Health (GA4GH), this subportal will disseminate access policies and procedures aligned to GA4GH using the Automatable Discovery and Access Matrix (or ADA-M: www.github.com/ga4gh/ADA-M). The Access Manager component will be provided as a simple user interface for researchers to apply for data access, as well as for cohort owners to facilitate the procedure of granting and managing granted credentials. For selected cohorts, the possibility of automatic credentials assignment based on applications and policies metadata will be explored through a blockchain technology developed by our SME member LYN in the MyHealthMyData EU project (www.myhealthmydata.eu).

3.3 Making data interoperable

Josep Gelpi, Jordi Rambla, Aad van del Lugt, ..Isabel Fortier?

harmonisation:Isabel Fortier euCanSHare will centralize data from several multi-site studies in a unique web-portal that will accelerate access to information and facilitate new data harmonisation.

integration: euCanSHare also will integrate highly heterogeneous data such as -omics and cardiac imaging. To do so, euCanSHare will exploit the expertise of all consortium members, to achieve the robust integration of all of these technologies.

- Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.
- Specify whether you will be using standard vocabulary for all data types present in your data set, to allow interdisciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?
- Project-Outline methodology for cross-referencing between repositories (ELIXIR/EGA and euro-BioImaging), and the technical solutions for enabling efficient access to data using the technologies developed by ELIXIR, EGA and euro-BioImaging.
- Project-Outline harmonization methodology

Q>
1-Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, 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)?
2-What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
3-Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?
4-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?

1 Interoperability of the tools will be maintained through a rich set of metadata allowing the system to associate tools and data in a transparent manner. Easy-to-use modules for authentication (based on KeyCloak) and secure data management (Oauth2 protocol for all encrypted data transfers) will be also integrated. The protocols to plug-in the tools (data browsers, visualisers, or analysis tools) on top of the main infrastructure will re-used from MuGVRE.Finally, execution scheduling will be based on a traditional queueing system to handle demand peaks in applications of fixed needs, and an elastic and multi-scale programming model (pyCOMPSs, controlled by the PMES scheduler) for complex workflows requiring distributed or multi-scale executions schemes. In euCanSHare, this will enable to derive tools execution to remote cloud infrastructures (through the OCCI protocol, www.occi-wg.org) and also to HPC environments within ELIXIR, EGA and euro-BioImaging.

Mica solution (www.obiba.org/pages/products/mica) developed by MUC for building the largest and most comprehensive easyto-use multi-cohort catalogue ever put together in the cardiovascular domain. This will help data custodians and network coordinators such as MORGAM, BiomarCaRE and CAHHM to efficiently organise and disseminate information about their cardiovascular studies and networks without significant technical effort.

EMC and MUHC will define metadata fields for the project, as collected by the euCanSHare cohorts and commonly used in cardiovascular research. Selection of the fields will be informed by existing standards adapted to serve the specific needs of the project. A Working Group will be established and convened at consensus meetings to define standard metadata for imaging, omics, epidemiological, clinical, and bio-sample data. EMC (euroBioImaging) will also develop and implement the models to support cardiac imaging metadata (imaging modalities, protocols, parameters and biomarkers), while MCGILL will focus on cataloguing cohort-specific access policies and consent requirements. Metadata fields will be populated with information obtained from participating cohorts (see Table 1 for the initial cohorts, while that new cohorts will be added through awareness campaigns). The final Mica-powered euCanSHare platform will include Maelstrom?s powerful search engine for allowing investigators to quickly find the information, variables and data they need for implementing cardiovascular research projects.

harmonization- from project proposal The project will leverage on the state-of-the-art technologies developed by the Maelstrom Research at MUHC (www.maelstromresearch.org), as well as the harmonisation models implemented during the MORGAM and BiomarCaRE projects (>40 cohorts harmonised). However, since it is impossible to perform a single harmonisation that will satisfy all future study requirements, we propose an original solution to facilitate future harmonisations by re-using previous harmonisation efforts in a more systematic manner. Specifically, we will store the harmonisation algorithms in a standardised electronic database such that any harmonisation effort can be easily searched and located in the database and re-used in new multi-cohort research studies, when relevant. With this approach, future harmonisations benefit from previous ones and new harmonisation rules/algorithms are stored to further populate euCanSHare?s harmonisation database. This approach will reduce cost and time of future multi-cohort research studies, while providing transparency on harmonisation processes. In this case, the harmonised dataset does not need to be stored; only the harmonisation rules and algorithms are saved in a standardised easy-to-search format and any harmonised data is generated on euCanSHare?s cloud by the software in real-time. In euCanSHare, the harmonisation database will be initially populated based on the BiomarCaRE and CAHHM harmonisation experiences, as well as based on use cases that will be investigated to test the platform. For implementing the proposed iterative harmonisation solution, MUHC?s software Opal20 will be adapted to provide a centralised web-based harmonisation management system allowing study coordinators and data managers to securely store/export a variety of data types and harmonisation rules in different formats using a point-and-click interface. Opal includes functionalities to define variables targeted for harmonisation, develop and implement processing algorithms used to derive common-format data, and efficiently document data harmonisation decisionmaking. To facilitate algorithm development, Opal also includes a comprehensive JavaScript library of functions commonly used to create harmonised variables. Establishing a secure connection with an R client also allows the use of the R programming language to derive common format variables. Opal then is executed to harmonise, store and display these data under the selected standardised model (e.g. BiomarCaRE?s or CAHHM?s model). Additionally, in euCanSHare, automated data quality control will be enabled through the Square 2 tool recently developed by UMG, which will enable to check for the level of normalisation, ambiguity and overall quality of both new sources slated for integration and the overall set of sources already integrated in the system. The resulting integrated harmonisation system will provide a highly flexible and efficient semi-automated process to on-board and harmonise new databases within the infrastructure.

3.4 Increase data re-use (through clarifying licenses)

- Specify how the data will be licensed to permit the widest reuse possible
- Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed
- Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why
- Describe data quality assurance processes

- Specify the length of time for which the data will remain re-usable

Q>
1-How will the data be licensed to permit the widest re-use possible?
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.
3-Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.
4-How long is it intended that the data remains re-usable? Are data quality assurance processes described?
3-In general, data collected in euCanSHare will become accessible through a centralized Web Portal and interoperable

interfaces integrating the tools from Table 2 in project proposal. This will be done by re-using the basis framework and IT solutions developed by BSC during the MuG project. Some data from some cohorts are restricted though. How to come about this. Every contributor should answer which part is restricted? 4-

4 Allocation of resources

- Estimate the costs for making your data FAIR. Describe how you intend to cover these costs.
- Clearly identify responsibilities for data management in your project.
- Describe costs and potential value of long term preservation

Q>
What are the costs for making data FAIR in your project?
How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).
Who will be responsible for data management in your project?
Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

5 Data security

- Address data recovery as well as secure storage and transfer of sensitive data

Q>
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?

6 Ethical aspects

- To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Q>
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).
Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

7 Other issues

- Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Q>
Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

[1] Paulus Kirchhof, Karin R. Sipido, Martin R. Cowie, Thomas Eschenhagen, Keith A A Fox, Hugo Katus, Stefan Schroeder, Heribert Schunkert, and Silvia Priori. The continuum of personalized cardiovascular medicine: A position paper of the european society of cardiology. *European Heart Journal*, 35(46):3250–3257, 12 2014.