PNRR – M4 - C2 - INVESTIMENTO 1.3

Decreto Direttoriale 341 del 15/03/2022 -

Avviso pubblico per la presentazione di Proposte di intervento per la creazione di "Partenariati estesi alle università, ai centri di ricerca, alle aziende per il finanziamento di progetti di ricerca di base"

Numero del progetto: PE_0000019

Titolo del progetto: HEALITALIA - Health Extended ALliance for Innovative Therapies, Advanced Lab-research, and Integrated Approaches of Precision Medicine

Area tematica: Diagnostica e terapie innovative nella medicina di precisione

Soggetto Attuatore: PALERMO - Università degli Studi

DATA MANAGEMENT PLAN

(DMP)

Document History

Version	Date (DD/MM/YYYY)	Created/Amended by	Changes
1	16/09/2022	Andrea Pace	-

Scheduled Data Management Plan (DMP) Updates

The DMP is a document that evolves during the lifespan of the project and registers all relevant changes in the life-cycle of all the research data sets of HEAL ITALIA. This document will be updated whenever important changes in the data or the data management policy occur.

The Data Management Plan (DMP)

The DMP will report the data management policies for all the data collected and generated by the HEAL ITALIA project, during and after its completion. Data will be described in detail for every single WP, task, and activity of within the project along with:

- the mechanisms in place for ensuring compliance with legal requirements
- the FAIR data strategy
- the Standard Operating Procedures
- the anonymisation/pseudo-anonymisation methods
- data curation, harmonization, and standardization methods
- data quality and data security management procedures

The DMP will include whether and how the data will be exploited and/or made accessible for verification and re-use.

The number, type, and nature of data and data sources may vary within the lifetime of the HEAL ITALIA project; the DMP will keep track of the actual data flow over time.

1. DATA SUMMARY

1.1 General Overview

Spoke 1: The main purpose of data generation and collection for SPOKE 1 is to create a mapping of the molecular environment of multifactorial diseases through a multi-omics analysis approach on biological matrices, aiming at identifying pathogenic genetic variants through the sequencing of DNA and refining “common-soil” hypothesis for the stratification of patients and the development of personalized therapeutical approaches through the analysis of the transcriptome. Moreover, proteomic analysis will be used to define the dynamic interactomes of the diseases and metabolomic alterations. Metabolomic analysis on mouse models or population cohorts, will allow the generation of metabolom maps. Altogether, the multi-omics approach will map the molecular landscape from the clinical to the molecular environment, to identify, classify, and refine the phenotypes of multifactorial diseases. Anonymised results will be monitored by the Spoke-1 International Scientific Board, by Patient Advocate, and released as open source.

Spoke 2: The main purpose of data collection and processing for SPOKE 2 is to design and develop a model for data curation, data quality management, and collaborative analytics among the partners of the HEAL ITALIA project, in full compliance with the Italian privacy legislation.

Spoke 3: SPOKE 3 sought to generate a model recapitulating the cellular complex evolution following the exposure to external agents (drug, radiation, etc.) for data generation and collection. The model will be corroborated by the use of 3D bioprinted organoids including fibroblasts, adipose, mesenchymal, endothelial cells or in vivo models. Machine Learning approaches, capable of reproducing and understanding the complex mechanisms of cellular communication, will be coupled with an integrated analysis of the 3D structures.

Spoke 4: The main purpose of data generation and collection for SPOKE 4 is to deliver new, cost-effective, evidence-based, predictive risk-based and non-invasive diagnostic and prognostic algorithm and pathways for accessible and affordable early detection, screening and progression risk-stratification of mono- and polygenic diseases and cancer. An integrated analysis of digital data including bioimaging, omics and data derived from medical devices, in-vitro and animal models, performed using computational tools is envisioned to define disease-specific biomarkers and personalized diagnostic, prognostic and therapeutic algorithms. Specifically, the addressed pathologies are rare diseases, metabolic diseases, cardiovascular diseases, and cancer.

Spoke 5: The main purpose of data generation/collection for SPOKE 5 is to create/design/develop innovative and comprehensive drug-screening and validation platforms allowing to overcome the limits of currently available systems and accelerate the identification of next-generation effective drugs in the field of precision medicine for oncology, rare diseases, microbiota alterations etc.

Our platforms are based upon the definition and implementation of a workflow composed by a series of intertwined in silico, *in vitro* and *in vivo* assays, tentatively grouped into successive stages of selection, progressively more refined and specific. A highly qualifying element of our approach is the use of unique human and murine *in vitro* and *in vivo* models, nano-technologies and high throughput screenings available thanks to the expertise gathered together in the present network.

Spoke 6: The main purpose of data generation and collection for SPOKE 6 is to design, develop and validate at the laboratory scale a portfolio of devices for addressing a given disease with biomedical tools, biomaterials and nanomaterials enabling a precise therapeutic approach, for monitoring a disease onset and progression and for monitoring the response to a therapy in ad hoc developed models. The data will be collected and analyzed with the main purpose of testing and validating such devices with test solutions and eventually with biological fluids together with cellular and small animal models when appropriate.

Spoke 7: The main purpose of data generation and collection for SPOKE 7 is to develop new composite tools able to predict with high accuracy the risk of disease onset or progression, to provide novel screening instruments for daily medical practice. The hallmark of SPOKE 7 will be to collect and integrate data on individual genetic features or biologic conditions in concurrence with environmental agents and life habits, with the final goal to improve existing disease prevention programs, thereby reducing morbidity and mortality related to the disease areas covered by the research project.

Spoke 8: The main purpose of data generation and collection for SPOKE 8 is to validate and implement, in the clinical setting, innovative predictive, preventive, diagnostic and therapeutic precision medicine approaches for people at risk of or affected by certain complex, multifactorial diseases, including some rare conditions. This will be founded on established or emerging molecular and clinical phenotyping and AI-driven decision-making protocols. Data will derive from research outputs coming from the other Spokes and also from a number of WPs of SPOKE 8.

1.2 Types of Data

Different types and data will be collected and processed in HealITALIA, depending on the aims of the specific Spoke and WP as detailed in the project workplan. In general, treated data will be including (but will not be limited to):

- Structured and unstructured clinical data (*e.g.*, Excel files, CSV files, SQL databases)
- Electronic Health Records (EHR) with both structured and unstructured data
- Diagnostic medical images (*e.g.*, DICOM, PEG, TIFF, GIF, PNG)
- High-throughput sequencing data (*e.g.,* FASTA, FASTQ, SAM, BAM, VCF)
- Digital scans of medical records (*e.g.*, PDF)
- Data streams from medical/wearable devices (*e.g.*, JSON, CSV, TXT)

More in general, Spokes will provide the data ecosystem for the other Spokes, collaborating in the identification of the most suitable standards and formats both for the data types produced by the spokes and their associated metadata, and providing adequate technical solutions for their adoption and management. Where appropriate, at the end of the project data, analysis methods, workflows, documentation, and software developed by the spoke will be released in trusted and well-suited public repositories with appropriate metadata included (*e.g.*, GitHub, GitLab, Zenodo, Galaxy Tool Shed) and/or preserved in designated servers of project partners with public access and exposing all the necessary metadata for data FAIRness.

1.3 Reuse existing data and methods used

To promote data reusability the data will be well-documented to support proper data interpretation, will have a clear and accessible data usage license so others know what kinds of reuse are permitted, will have provenance information to make clear how, why and by whom the data have been created and processed, will meet relevant domain standards. A clear license to govern the terms of data reuse will be provided such as Creative Commons (CC).

In general, retrospective datasets have multiple data sources that provide for more joint data controllers or a single data controller. Partners transmitting/receiving/processing data will sign a data processor agreement. Data will only be transmitted/received/processed after verification that the use/reuse is in accordance with EU and Italian Privacy legislation. Appropriate anonymization or pseudo anonymization methods will be applied by data controllers. Following FAIR principle of data management (section 2), data collected and generated within the project will be curated, standardized, and harmonized adopting, whenever possible, standard data models, e.g. the Observational Medical Outcomes Partnership (OMOP) common data model, and standard vocabularies, such as ICD, SNOMED, and LOINC. The coding of the variables and the transport/document standards will be identified and approved by the Parties. Methods, software and workflows developed for data management and standardization will also be inherently re-usable beyond their primary purpose by applying them on similar data produced in other contexts.

Moreover, data collected and generated can be used for academic purposes, such as: scientific dissemination, training and education programs, meetings and/or publications; communication of relevant data to Regulatory Agencies for authorization. Each data collection and processing activity within HEAL ITALIA will undergo an assessment process that will establish the most appropriate use and reuse licensing models, chain of responsibility and conditions for data processing, and authorized uses.

1.3 Dataset reference and description

Data/metadata will be assigned a globally unique and persistent identifier (repository unique identifier).

By M6 all the SPOKEs will release a detailed description of the retrospective and prospective data sources together with the expected data flow and the adopted SOPs.

The categories of data considered within the project are:

- *Raw collected data* – both prospective and retrospective data not yet subjected to quality assurance or control.
- *Validated collected data* – raw data that have been evaluated for completeness, verified for compliance with standard operating procedure and validated for specific quality.
- *Derived data* – data which have been generated through statistical operations and/or through qualitative and quantitative analysis and processing of Raw and Validated data

1.5 Expected Dataset Size

Actual size of Data gathered will be calculated during project run and it will be powered at Spoke level. The expected size of the envisioned datasets is an educated guess based on an estimation of the average data size and number of measurements requested, but it may vary with respect to what is declared in the present document. The dimension of total collected data will be in the order of magnitudes of hundreds of terabytes.

2. FAIR DATA

2.1 Making Data Findable, including provisions for metadata

To make the data used by each spoke Findable, guidelines will be provided on the use of machine-readable standard metadata for automatic discovery of datasets and services.

Research data will be organized in data sets, which are named collections of data units with the same focus and scope. A standard naming convention will be used for files during data collection and handling and the version number of the data set will be added at the end of the title in case of data revisions, to help identifying updates. Data/metadata generated will be registered and indexed in searchable sources, and data will be described with rich metadata (the dataset's context, quality, condition and characteristics will be described).

Whenever the nature of the data would allow it, data produced during the project activities will be deposited and described in institutional or public data repositories. All documents or data uploaded

to repositories will be granted a Persistent Identifier (PID, e.g., a Digital Object Identifier, DOI). The chosen repositories will also support the use of standard descriptive metadata to grant proper description and documentation of the data. Involved researchers will be registered at ORCID external link for persistent author identifier and this identifier will be used with all (data) publications.

Due to the sensitive nature of much of the data that will be produced during the project activities, we envision that only fully anonymized data will be made publicly available (e.g., aggregated data) while for the other datasets only metadata can be made publicly available.

All data regarding human subjects will be anonymized or pseudo-anonymised in accordance with current regulations. Search keywords will be provided to optimize both findability and re-usability. Dataset specific keywords will also be included describing the property characterized within the dataset and the method or instrumentation used to collect/produce the data. Open access repositories or equivalent will be adopted, where applicable, for the sharing of Data and metadata. At the moment of publication of project results, each research team and Spoke leaders will verify the fulfillment of legal requirements for making the relative underlying data publicly available. Whenever possible, the data will be made available through trusted data repositories.

2.2 Making Data Accessible

Considering the variable nature of the data that will be produced in the HEAL ITALIA project, each use case has specific needs and policies for what concerns open access to data. In general, the project will follow the FAIR data principles – Findable, Accessible, Interoperable and Reusable. It has to be noted, though, that FAIR does not necessarily imply that data will be openly available in all cases, as implied by the paradigm "As Open as Possible, as Closed as Necessary". Specifically, restrictions on data access or impossibility to share them will be considered only in the following cases:

- when collected data belongs to third party which have denied permission for sharing them on account of confidentiality and proprietary issues;
- protection of personal data of key informants involved in surveys, focus groups, interviews, and case studies;
- when availability of the data would mean that the project's main aim might not be achieved.

Sensitive or IP protected data cannot be made available unless via anonymization / obfuscation procedures. For data that fall within some of the limitations described above, and for which no action can be taken to make them shareable, prevented or restricted access will be considered. Further versions of the DMP will indicate the versions or parts of the datasets that cannot be freely shared, giving specific reasons, and will further describe the procedures adopted by the consortium in case of access limitations.

Physical access in data centers will be subject to multiple layers of access control (at least two levels of access control). Remote login access is password protected and access rights will only be granted by Spoke Leaders to authorized individuals. All information will be coded, and participants will be assigned unique study numbers. An investigator at a particular site will only access data from their own patients. The Spoke Leaders and designated authorized members of the data management team may access all anonymised data with authorisation granted by the Spoke Leaders. The Spoke Leaders make a commitment to maintaining the confidentiality, safety, security, and integrity of all

confidential and sensitive data, which is held under its guardianship. Anonymised data may be shared with the study team and their affiliates in order to meet the objectives of the study.

Each Spoke leader will survey members to understand if Data is already available in a trusted repository and whether it's appropriate to openly share them. Many partners have data which are much broader than those required for the Project. Thus, they may need time to fully explore the implications of depositing data in a trusted repository and/or to generate a reduced dataset containing only variables relevant to this project. Where possible, each Spoke will follow open-access rules provided by European standards such as those of Horizon Europe. Open data can be used by third parties, possibly in different contexts, to generate new beneficial results, including new open data. Only for protected data, a formal application shall be submitted, countersigned by the legal representative of the researcher's Institution. The genetic/medical/imaging data will be embargoed until publication (on a preprint archive or a scientific journal). Spokes will provide open access (OA) to research outputs (*e.g.*, publications, data, software, models, algorithms, and workflows) through deposition in trusted repositories. In fact, partners will provide OA to peer-reviewed scientific publications relating to their results. The authors of all peer-reviewed scientific publications will choose the most appropriate way of publishing their results, and these publications will be stored in an OA trusted repository, during and after the project's life.

2.3 Making Data Interoperable

The consortium will strive to collect and document the Data in a standardized way to ensure that datasets can be easily used by researchers and institutions in different countries. Each Spoke leader will survey members and data providers in general to carry out a complete inventory of the sources of information, their format, their coding, and the legal basis for using the data for the purposes of the project. Where non-standard vocabulary will be necessary, metadata will include detailed documentation about the vocabulary and general methodologies employed for the generation of the dataset. The inventory and related SOPs will be released by the Spoke leaders by M6.

A dedicated file (.txt or .pdf) will be published to help ensure that data can be correctly interpreted and re-analysed by others. The dedicated data file will contain the following information:

1. for each filename, a short description of what data it includes, optionally describing the relationship to the tables, figures, or sections within the accompanying publication;
2. for tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units;
3. any data processing steps, especially if not described in the publication, that may affect interpretation of results;
4. a description of what associated datasets are stored elsewhere, if applicable;
5. whom to contact with questions

Common standards, practices, infrastructures, and governance framework will be selected in accordance with European guidelines to ensure interoperability within as well as outside the consortia.

2.4 Increase Data Re-use

Considering the nature of the data obtained in the project, and the research in the precision medicine area, all the Spokes will follow the best practices for scientific data publication and sharing with the appropriated public domain databases and data repositories. This is highly dependent on the spoke / use case. Whenever the nature of the data would allow it, data produced during the project activities will be deposited and described in institutional or public data repositories. All documents or data uploaded to repositories will be granted a Persistent Identifier (PID, e.g. a Digital Object Identifier, DOI). Each partner will self-evaluate which background information will be accessible without any restriction to other researchers. To avoid any potential doubt, the consortium will attach specific licenses to the deposited data to define all conditions under which the work is provided under either open or restricted access. Some data and metadata will be embargoed until publication (on a preprint archive or a scientific journal). Long time preservation of data will be ensured by the selected data repositories, which will be chosen taking into account their specific preservation policies.

Backup copies of datasets that cannot be publicly distributed will be kept by data controllers and data processors responsible for their collection and management in project infrastructures compliant with the specific legal requirements. Depending on the spoke and on the nature of the produced dataset, project participants will distribute the shareable data by adopting licenses that allow re-use of the data and of the data sets by other scholars and stakeholders. The data sets will be made available, unless otherwise stated, under the latest available version of the Creative Commons Attribution International Public License (CC BY) or Creative Commons Public Domain Dedication (CC 0) or a license with equivalent rights, following the principle ‘as open as possible as closed as necessary’.

3. RESOURCE AND ALLOCATION

3.1 Costs for FAIR Data (and long-term retention of data)

The non-identifiable data will be retained by the HEAL ITALIA consortium for 5 years after the final publication from the studies. Where applicable, storage costs are accounted for by funds allocated to each partner for storage resources and by each partner allocating personnel to curating the data. The final datasets will be available for sharing with other researchers 12 months after publication of the study results and will be shared in public repositories. After closure of study sites, the data will be archived in the public repository, as previously specified.

The activities related to making the data/outputs FAIR are anticipated to be covered within the allocated budget for each spoke. The estimation of such cost is not possible at the moment because it depends on various elements, among which: 1) the fees applied by chosen repositories for long term archiving and data curation; 2) the storage solution used among partners to share data; 3) the tools of election for sensitive data handling (e.g. for pseudonymization/anonymization, encryption); 4) the evaluation of costs related to data management and documentation, conversion of files into open formats, deposit (also in terms of Person-Month). Next versions of the DMP will detail strategies and associated costs for long-term data hosting.

3.2 Responsible for Data Management

Each Data controller and Data processor will be responsible for data management at the beneficiary institution, supported by the institution Data Manager and/or Data Protection Officer. The Project Manager will coordinate and monitor the activities for ensuring compliance to the project DMP.

4. DATA QUALITY

Data quality assurance will be performed by all partners to ensure data inconsistency, missing information, and application of wrong data analysis methods are considered and acted upon, to guarantee data precision and accuracy. Tools to conduct quality checks (such as OpenRefine), will be decided upon project start and applied by all partners. Only data passing stringent QC will be disseminated. In this regard, it is important to underline how data scientists and data engineers will monitor the processes to identify errors using a statistical approach. A special effort will be spent to ensure Accuracy, Relevancy, Completeness and Timeliness.

5. DATA SECURITY

Each partner will have to process sensitive data in some form, ranging from service applications data (*e.g.*, names, affiliations, and emails from users) to aggregated research data (*e.g.*, genomes, biomarkers). Data generated by each partner (primary, derived, and interpreted data) will initially be stored at local level, in local GDPR-compliant infrastructures. Spoke and WP leaders will perform an extensive risk analysis and will adopt appropriate risk avoidance and mitigation technical (e.g., data backups, secure networks, and firewalls) and organizational (e.g., access right management, operating instructions, and staff training) measures. Following the European and national rules and regulations regarding personal data protection, it is always required to define minimum data retention time. Any processing of personal data will respect privacy by design and by default principles and will happen in full compliance with regulations in force and following the principles of necessity, correctness, pertinence, and minimisation, data processing takes place solely for determined, explicit, and legitimate purposes.

The following measures will be put in place to comply with GDPR requirements for the safeguard of sensitive personal data: 1) data encryption (policies will be established for the creation and distribution of secure encryption keys as well as for the access to the data itself.); 2) controlled access registries and systems for data access management (REMS or equivalent); 3) pseudonymization and de-identification of primary data; 4) if/when applicable data will be deposited in controlled access repositories recognized by the international scientific community, data access will be governed by a dedicated Data Access Committee; 5) Redundant storage systems for preserving data integrity over time. Data security will be assessed and evaluated based on:

- *Integrity*: measures will be taken to ensure that data remain intact and unaltered as audit trails and change control.
- *Accessibility*: to prevent data loss, cloud backup service will be implemented to complement local storage.

- *Confidentiality*: measures will be implemented to protect the confidentiality of sensitive data as restricted access, anonymization and pseudonymization.

Additional key data protection principles, as defined by the GDPR, include lawfulness, fairness and transparency, purpose limitation, data minimization, storage limitation and measures to ensure accountability.

6. ETHICAL ASPECTS

All activities within HealITALIA will comply with all necessary legal requirements and ethical principles applicable under international, EU and national law, the ethics provisions set out in the Grant Agreement, and follow the highest ethical standards. The partners adhere to current ethics legislation and regulations in the countries where their research will be carried out relating both to ethics and data protection issues. All research that involves humans and human data is conducted under the rules and legislation in place, including the Declaration of Helsinki, the IHC guideline for Good Clinical Practice, as well as the ISO guidelines on good clinical practice ISO 14155:2020, the European Directive 2001/20/EC on Good Clinical Practice, and the EU General Data Protection Regulation (2016/679). State of the art and up-to-date compliance measures will be employed ensuring correct conduct under the legal and ethical guidance of the GDPR and applicable national laws concerning patient data. Data privacy and confidentiality will be provided based on anonymisation techniques, security-by-design, and data-protection-by-default principles (e.g. persistent data anonymization and abstraction in inter-system data transfer and interoperability, system and platform requirements, but also human-in-the-loop interactions). Attestation techniques (CFA) will enhance the framework's operation by addressing edge security vulnerabilities, preserving data integrity, and most importantly creating a network of trust. To ensure the transparency and trustworthiness of the Artificial Intelligence systems and methodologies, all activities concerning the design, development and deployment of these models will adhere to Ethics Guidelines for trustworthy AI, as formulated by the High-Level Expert Group on Artificial Intelligence (AI HLEG).
Any research activity involving humans or animals must receive the approval of an official Research Ethics Committee.

Mouse Studies

Experiments will be performed after approval by the national ethical committees for animal welfare and under authorization through the appropriate national laws and regulations, in compliance with the EU Directive 2010/63/EU and its implementations in national legislation (Decreto Legislativo del 4 Marzo 2014 n° 26).

In compliance with the 3R principle (reduction, refinement and replacement), whenever possible in vitro cell culture studies will be performed to substantially replace in vivo studies. Pilot in vivo

experiments with 2-3 animals/group will be run. Optimum number of animals will be defined via i) cost/benefit analysis and ii) statistical analysis to ensure an acceptable level of confidence in any particular read-out or end-point of the experiment.

Personnel involved in the maintenance of mouse colonies and experimentation will either be professionally employed by the animal house or fully trained (official certification as per national legislation) before experiments are conducted. Mice will be kept on chow diet and appropriate measures will be taken to minimize animal suffering (e.g. anesthesia before killing, daily inspection during treatment for immune cell depletion) and to protect their welfare (e.g. small number of animals/cage and appropriate environmental enrichment).

All researchers are required to follow this general strategy to ensure compliance with International standards on ethics and data management.

Moreover, Spokes will consider at all steps of the project all possible additional ethical issues arising in the project course, thanks to experts in safety and ethical issues who are part of the consortium.

F.to Prof. Andrea Pace
Presidente della Fondazione Heal Italia
Prorettore alla Ricerca, al Trasferimento Tecnologico
e ai Rapporti con l'Amministrazione
dell'Università degli Studi di Palermo

Firmato Digitalmente in Pades
Ai sensi del codice sull'Amministrazione Digitale