MELCAYA

NOVEL HEALTH CARE STRATEGIES FOR MELANOMA IN CHILDREN, ADOLESCENETS AND YOUNG ADULTS

Grant Agreement: 101096667

D10.3 Data management plan 1

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Document Information

Deliverable number:	D10.3
Deliverable title:	Data management plan 1
Deliverable version:	v1.0
Work Package number:	WP10
Work Package title:	Project management
Due Date of delivery:	31.05.2023
Actual date of delivery:	31.05.2023
Dissemination level:	Public (PU)
Type:	R - Document, Report
Author(s):	Adrián López (FCRB)
Contributor(s):	All partners Silke Fiers (LUCIA)
Reviewer(s):	Dario Sacchini (UCSC) Pietro Refolo (UCSC) Costanza Raimondi (UCSC) Mario Mandala (UNIPG) Laura Sampietro (HCB) Gidi Shani (TECH) Tabea Bucher (DKFZ) Josep Maria Borràs (ICO) Silke Fiers (LUCIA)
Project name:	Novel health care strategies for melanoma in children, adolescents and young adults
Project Acronym:	MELCAYA
Project starting date:	01.12.2022
Project duration:	48 months
Rights:	MELCAYA consortium

Document history

Version	Date	Beneficiary	Description
0.1	11.05.2023	FCRB	First draft version
0.2	12.05.2023	UCSC	Revised draft
0.3	12.05.2023	UNIPG	Revised draft
0.4	15.05.2023	HCB	Revised draft
0.5	16.05.2023	TECH	Revised draft
0.6	16.05.2023	DKFZ	Revised draft
0.7	19.05.2023	ICO	Revised draft
0.8	21.05.2023	LUCIA	Addition of the common chapter for the *Understanding* cluster
1.0	26.05.2023	FCRB	Final version

Executive Summary

The aim of this deliverable is to outline the data management strategy for MELCAYA project, including a description of the actions and processes related to the collection and management of data during the project. The first part of the document reviews the purposes of data collection, as well as the overall architecture for the flow of data and presents a summary of the preliminary datasets to be used throughout the project (including aspects such as the type of data, format, expected size, etc). The strategies to be implemented to ensure that all data produced throughout the project remains findable, accessible, interoperable and reusable (FAIR principles) are also presented, including provisions on permanent identifiers, metadata, repositories and databases to deposit data, etc. A description regarding the allocation of resources to guarantee continued accessibility and data discovery after project completion, specific responsibilities for the data management activities and the general security measures for data protection is also presented. Besides this information, an overview of the main ethical and legal aspects related to data management is presented, including a description of the legal and ethical basis for data processing, the procedures for data pseudoanonymization and how incidental or secondary findings would be communicated to the participants in the clinical studies. Finally, a brief description of the common areas for collaboration in terms of data management with other projects of the *Understanding cancer* cluster is presented. This report provides a preliminary data management plan that will be continuously updated as the project progresses, with more consolidated revisions being scheduled for months 24 and 36.

Contents

Figures

Tables

Acronyms & Abbreviations

Term	Description
EU	European Union
EC	European Commission
DMP	Data Management Plan
WP	Work Package
M	Month
EMA	European Medicines Agency
AI	Artificial Intelligence
GDPR	General Data Protection Regulation
DoA	Description of Action
GA	Grant Agreement
CA	Consortium Agreement
IP	Intellectual Property
PC	Project Coordinator
EMB	Ethics Monitoring Board
EC	Exploitation Committee
JCA	Joint Controller Agreement
MTA	Material Transfer Agreement
CAYA	Children, Adolescent and Young Adults
eCRF	Electronic Case Report Form
N/A	Not applicable
RNA	RiboNucleic Acid
DNA	DeoxyriboNucleic Acid
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing
SNP	Single Nucleotide Polymorphism
CGH	Comparative Genomic Hybridization
FISH	Fluorescence In Situ Hybridization
ENA	European Nucleotide Archive
EMBL-EBI	European Bioinformatics Institute

EGA	European Genome-phenome Archive
WSI	Whole Slide Images
H&E	Haematoxylin/Eosin
FFPE	Formalin-Fixed Paraffin-Embedded
MICS	MACSima Imaging Cyclic Staining
AJCC	American Joint Committee on Cancer
DOI	Digital Object Identifier
ISIC	International Skin Imaging Collaboration
GEO	Gene Expression Omnibus
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography

1 Introduction

The goal of this document is to describe the data management life cycle for the data collected, processed, generated and published in MELCAYA. It will provide an overview of all datasets to be generated by the project, define the management policy to ensure that all data produced respects FAIR principles (findable, accessible, interoperable and re-usable) and related ethical aspects such as the communication of incidental or secondary findings. A special section will be devoted to the data management within the *Understanding* cluster, in which commonalities in data standards, validation, protection and data exchange will be commented. This document will follow the reference structure for the Data Management Plan (DMP) proposed by the European Comission (EC) for the *Horizon Europe* framework programme. As we are quite early in the project, with the different clinical study protocols still under development, it is not possible to provide a complete description of all datasets and related details. Final protocols and ethics approvals need to be ready by M11 (October 2023) with next versions of the DMP scheduled for M24 (November 2024) and M36 (November 2025).

2 Data summary

2.1 Purpose of data collection and relation with the objectives of the project

The main goal of MELCAYA project is to obtain insights into the mechanisms of melanoma development and progression in children, adolescent and young adults under 30 years old (CAYA) to create new prevention strategies and develop innovative technologies for diagnosis and prognosis of the disease. To achieve this overarching goal, several sub-objectives will have to be met by the implementation of 7 different clinical studies:

- Identification of the environmental and genetic risk factors for melanoma in CAYA (*study 1: ExpoMel*).
- Identification of the molecular profiles of progression from benign congenital nevi to melanoma (*study 2: NevustoMel*).
- Development and promotion of international protocols and standard procedures for melanoma taxonomy (*study 3: MolMel*).
- Evaluation of the clinical efficacy and safety of anti-PD1 antibodies in CAYA melanoma patients (*study 4 and 5: ImmunoPed I and II*).
- Development of artificial intelligence (AI)-based diagnostic tools to distinguish images of melanomas from images of nevi or other benign pigmented lesions (*study 6: AI-Mel*).

- Development of novel rapid and non-invasive tools for early detection, risk stratification and prognosis of melanoma in CAYA (*study 7: PrecisMel*).

These clinical studies will require the collection and processing of different types of data, which can be grouped in the following general categories:

- **Clinical data:** including variables such as age, sex, anatomical site of skin lesion, previous diseases and treatments, tumor stage, etc.
- **Epidemiological data:** incidence of melanoma in different countries.
- **Exposomic data:** including variables such as air pollution, climate factors, UV exposure, diet, socioeconomic status, etc.
- **Genomic data:** including information such as germline and somatic mutations, copy number variants, fusions, etc. obtained using RNA sequencing, whole genome sequencing or whole exosome sequencing techniques on biopsy tissue or blood samples.
- **Medical images:** including dermatoscopic and histopathological images.
- **Volatilomic data:** electrical resistance profiles generated by different volatile organic compounds (VOCs) obtained with the breath analyzer and disposable sensing patch manufactured at the Israel Institute of Technology (TECH).

2.2 Overall data management

The preliminary idea is that each of the participating institutions contributing to the clinical studies upload their data to a secured electronic Case Report Form (eCRF) system based on REDCap [1] and hosted by coordinating partners Fundación Clínic per a la Recerca Biomèdica-Hospital Clínic de Barcelona (FCRB-HCB). RedCap complies with international standards on data protection and offers a consistent, auditable and integrated electronic database environment. The structure of this eCRF will have a core module with general fields regarding patient (sex, age, previous diseases, etc.) and melanoma(s) information (anatomical site, tumor stage, date of relapse/progression, etc) and then extensions that would stem from it for the different work packages (see figure 1). Access to the platform will be password protected and electronic login credentials will be issued only to named authorized individuals (investigators, monitors, data managers).

For instance, for WP1 particular fields would be available to upload information such as ZIP code of the patient and related exposomic data for that area (air pollution, UV exposure, etc.) as well as information on germlike mutations. On the other hand, for WP4 other fields would be available such as reported toxic effects, efficacy, effects on fertility, etc. All available medical information about a

patient will not be uploaded in its entirety to the server, rather, only selected data fields that are relevant and necessary for the project in a completely pseudoanonymised format (as explained in section 6.2) so that the primary identifier of the patients won't be available at all. For those partners not willing to upload patient information directly to REDCap, it will be possible to export the database structures and use them offline.

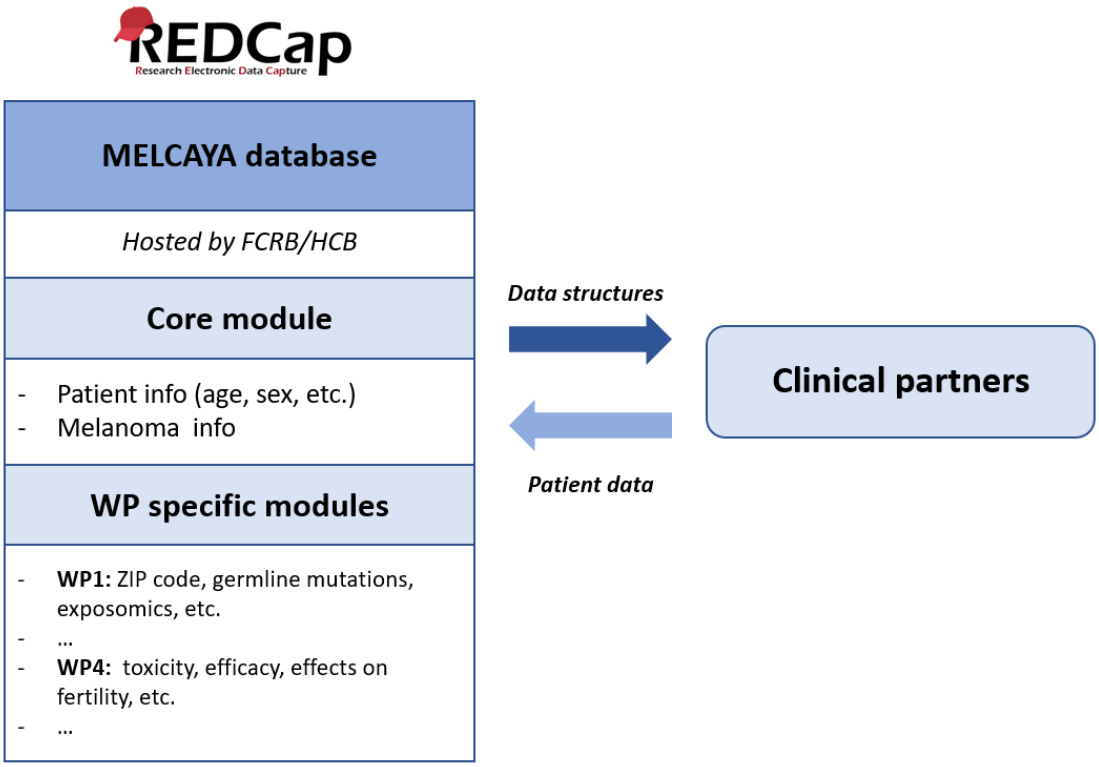

Figure 1 Preliminary data management workflow for MELCAYA

2.3 Overview of data to be processed

In this first iteration of the DMP there are still many unknown details about data that will be clarified as the project progresses. A preliminary identification of the main datasets to be processed during MELCAYA with our current understanding and how this data will be collected and stored has been performed among consortium partners by answering the questions defined in the *Horizon Europe* DMP template (see tables 1 to 10):

Table 1 Data summary for clinical data

What type of data is generated?	Clinical information about the patients including age, sex, previous diseases, stage, tumour burden, previous treatments, concomitant diseases and medications, scheduled therapy,

	clinical response evaluation in solid tumours, date of relapse/progression, date of death, efficacy, adverse effects (including fertility) and molecular features.
What is the origin of the data? Is existing data going to be re-used?	Most of the data will be re-used from available databases of clinical partner institutions and international cancer registries such Catalonia Melanoma Network (Xarxa Melanoma), the German Registry for Rare Pediatric Tumours (STEP), the Swedish Melanoma Registry (SweMR), the Public Pathology Database of the Netherlands, the melanoma collection from the European Organization for Research and Treatment of Cancer (EORTC), the Italian Oncology Paediatric Network under the auspices of AIEOP, the Italian Melanoma Intergroup (IMI), the Polish Melanoma Academy as well as selected hospitals from the French Society of Paediatric Oncology. Prospective patient recruitment is also envisioned in the *Immuno-Ped study II* (WP4) and in *Precis-Mel* study (WP6).
What is its format?	PostgreSQL, CSV, JSON, XLS
What is the overall expected size of the data?	2 GB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP1 to 6
What will be the utility of the data?	Clinical data will be needed for the interpretation of the molecular and volatilomics data as well as for the evaluation of the efficacy and adverse effects (including fertility) of anti-PD1 therapy in CAYA.

How is data planned to be stored?	Data will be stored locally at each partner's location and transferred to the common shared database in RedCap hosted by FCRB/HCB in a pseudoanonymised format.

Table 2 Data summary for genomic, transcriptomic and epigenomic data

What type of data is generated?	Data generated by RNA sequencing (RNAseq), whole genome sequencing (WGS), whole exosome sequencing (WES), single nucleotide polymorphism (SNP), comparative genomic hybridization (CGH), fluorescence *in situ* hybridization (FISH) and methylomics.
What is the origin of the data? Is existing data going to be re-used?	Data re-used or obtained from previously collected blood and tumor tissue samples stored in different registries and institutional biobanks from different clinical partners such as the Catalonia Melanoma Network (Xarxa Melanoma), the German Registry for Rare Pediatric Tumours (STEP), the Swedish Melanoma Registry (SweMR), the Public Pathology Database of the Netherlands or the melanoma collection from the European Organization for Research and Treatment of Cancer (EORTC).
What is its format?	*CSV, FASTA/FASTQ, BAM, VCF* and *XML*
What is the overall expected size of the data?	5 TB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	Mainly WP1 and WP2
What will be the utility of the data?	Germline sequencing will be used for the identification of melanoma predisposition in our patients and calculate a risk score in combination with exposomic data. Tumour

	sequencing data will be used to support and complement the germline results. Tumour sequencing will also serve to differentiate between different entities: for instance, the assessment of copy number variations would allow for the differentiation between benign proliferating nodules and melanomas on congenital nevi.
How is data planned to be stored?	Data will be stored locally at each partner's location and transferred to the common shared database in RedCap hosted by FCRB/HCB. Results of the analyses will be shared (depending on previous informed consent forms) on public databases such as the European Nucleotide Archive (ENA) at the European Bioinformatics Institute (EMBL-EBI) or the European Genome-phenome Archive (EGA).

Table 3 Data summary for histopathological images

What type of data is generated?	Digitalized histopathological whole slide images (WSI).
What is the origin of the data? Is existing data going to be re-used?	Re-use of images produced from the haematoxylin and eosin (H&E) staining of formalin-fixed and paraffin-embedded (FFPE) tumor tissues and frozen tumor tissue samples from the institutional biobanks of the different clinical partners.
What is its format?	*SVS*
What is the overall expected size of the data?	1.5 TB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP3 and WP5

What will be the utility of the data?	To develop a new hybrid taxonomy for tumor sample classification by combining morpho-molecular and immunophenotypic data with somatic, transcriptional and epigenomic data and to train a machine learning model for the early diagnosis of melanoma and distinction from nevi and other benign lesions.
How is data planned to be stored?	Data will be stored locally at each partner's location and transferred to the common shared database in RedCap hosted by FCRB/HCB.

Table 4 Data summary for spatial proteomics

What type of data is generated?	High-content immunofluorescence imaging
What is the origin of the data? Is existing data going to be re-used?	New images will be obtained using an automated ultra-high content imaging technology (MICS) for the simultaneous analysis of hundreds of marker antigens on previously collected samples available in the institutional biobanks of the different clinical partners.
What is its format?	*TIFF, PNG, JPEG*
What is the overall expected size of the data?	1 TB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP2 and WP3
What will be the utility of the data?	Discovering novel diagnostic signatures of tumor cell phenotypes and providing a tissue-based protein atlas according to parameters such as morphology, multi-parameter cell lineage assessment (spitz vs. spitzoid) as well

	as the location of individual tumor and immune cells subsets within the tumor microenvironment.
How is data planned to be stored?	Data will be stored locally at each partner's location (mainly UT) and transferred to the common shared database in RedCap hosted by FCRB/HCB.

Table 5 Data summary for dermatoscopic images

What type of data is generated?	Dermatoscopic images of melanoma and nevi
What is the origin of the data? Is existing data going to be re-used?	Mostly re-used images obtained with a an epiluminiscence microscope (dermatoscope) available in the data repositories of different clinical partners. Some images will also be prospectively collected in the *AI-Mel* study (WP5).
What is its format?	*PNG*
What is the overall expected size of the data?	10 GB
What metadata is used to describe the data?	Primary diagnosis (with subtypes), pigmentation, American Joint Committe on Cancer (AJCC) stage, Clark level, Breslow thickness, maximal diameter, mitosis rate, ulceration, approximated age, gender, localization and Fitzpatrick skin type.
What WP(s) will make use of this data?	WP5
What will be the utility of the data?	Training of a machine learning-based image classifier for the early detection of melanoma lesions in CAYA and distinction from benign lesions to help dermatologists during routine skin cancer screening.

How is data planned to be stored?	Data will be stored locally at each partner's location (mainly DKFZ) and transferred to the common shared database in RedCap hosted by FCRB/HCB.

Table 6 Data summary for exposomic data

What type of data is generated?	Geographical data, climate variables (such as surface temperature, solar radiation, wind speed, etc.) and air pollutants (PM10, PM2.5, NOx, O_3, SO_2, CO, etc.)
What is the origin of the data? Is existing data going to be re-used?	Re-use of data from ground-based instruments (such as air quality monitoring stations, sum photometers, etc.), satellites (Terra MODIS, Aura OMI, Sentinel TROPOMI) and re-analysis of data from Copernicus Atmosphere Monitoring Service (CAMS) as well as government/administrative data.
What is its format?	*XLS*
What is the overall expected size of the data?	10 GB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP1
What will be the utility of the data?	Identification of relevant risk factors for melanoma in CAYA and creation of a risk score in combination with genomic data.
How is data planned to be stored?	Data will be stored locally at IUF and transferred to the common shared database in RedCap hosted by FCRB/HCB.

Table 7 Data summary for epidemiological data

What type of data is generated?	Epidemiological data on the incidence of melanoma in CAYA

What is the origin of the data? Is existing data going to be re-used?	Re-use of data retrieved via international databases such as the Global Cancer Observatory (GLOBOCAN) and EUROCARE. National databases like the German Childhood Cancer Registry will also be used.
What is its format?	*XLS*
What is the overall expected size of the data?	1 GB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP1
What will be the utility of the data?	Understanding differences in melanoma incidence between countries and identification of relevant exposomic risk factors for melanoma in CAYA.
How is data planned to be stored?	Data will not be stored.

Table 8 Data summary for volatilomic data

What type of data is generated?	Electrical resistance measurements caused by volatile organic compounds (volatilomics)
What is the origin of the data? Is existing data going to be re-used?	Prospective data obtained from recruited patients at HCB using the disposable sensing patch and the breath analyser device manufactured by TECH.
What is its format?	*CSV, JSON*
What is the overall expected size of the data?	2 GB

What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP6
What will be the utility of the data?	To develop a screening tool for early detection and diagnosis of metastasis in CAYA patients
How is data planned to be stored?	Data will be stored in the common shared database in RedCap hosted by FCRB/HCB.

Table 9 Data summary of related projects and publications

What type of data is generated?	Collection of related projects, papers, publications, reports, books, press releases, newsletters, etc.
What is the origin of the data? Is existing data going to be re-used?	Re-use of publicly available data as well as new data generated throughout the project through publications, press releases, etc.
What is its format?	*DOC, PDF, MP4*
What is the overall expected size of the data?	1 GB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP1 to W10
What will be the utility of the data?	The data will be used by consortium partners and external organisations for networking, awareness raising purposes, spreading and building upon knowledge about pressing issues and challenges in the field as well as to generate possible solution approaches.

How is data planned to be stored?	The collected data will be stored in the project's OneDrive folder and made accessible through public project reports as well as through articles on the project website.

Table 10 Data summary of workshops, consultations and focus groups

What type of data is generated?	Virtual consultations, workshops, group interviews (DELPHI method) and focus group data.
What is the origin of the data? Is existing data going to be re-used?	New data generated throughout the project during discussions of panels of experts on Ethics, Sociology and Law, workshops with patients and through the creation of a forum with EU stakeholders, including patient organizations, scientific and medical organizations, local and European government agents, health technology agents, payers and regulatory bodies.
What is its format?	*XLS*
What is the overall expected size of the data?	500 MB
What metadata is used to describe the data?	N/A
What WP(s) will make use of this data?	WP7 to WP9
What will be the utility of the data?	The data will be used to obtain new insights about health strategies and technologies for the prevention, screening and early detection of childhood melanoma, as well as the associated ethical, social and patient perspective. It will also be used for the education of patient and patient advocates and to ensure the involvement and engagement of the wider European melanoma community.

How is data planned to be stored?	The collected data will be stored in the project's OneDrive folder and made accessible through public project reports as well as articles on the project website.

3 FAIR data

The research outputs generated during MELCAYA project will follow FAIR principles [2]: findability, accessibility, interoperability and re-usability. They describe how to organize the produced data so they are more easily accessed, understood, exchanged and re-used. These have a key importance for major funding bodies such as the EC as they allow to maximize the integrity and impact of the research investment. The strategies to implement FAIR principles in this project are the following:

3.1 Findability

Data generated throughout the project lifetime will facilitate scientific discovery by providing an important research resource for the field of genomics, exposomics and machine learning in medicine. To encourage rapid availability of key findings and promote data sharing, unique and persistent identifiers such as DataCite and Digital Object Identifiers (DOI) will be used in order to link MELCAYA datasets with scientific publications and other research outputs.

3.2 Accessibility

Data will be deposited in open repositories according to each institution's open access policy. Generalist data repositories within the Open Access Infrastructure for Research in Europe, such as OpenAIRE [3] or Zenodo [4] will be used in order to ensure the widest degree of public access possible. Clinical and dermatoscopy images of tumors will be deposited in public archives as ISIC (International Skin Imaging Collaboration) [5] and the European Cancer Imaging Initiative [6]. Genetic and transcriptomic data will be deposited in a suitable discipline-specific repository such as the European Genome-phenome archive (EGA) [7] or the Gene Expression Omnibus (GEO) repository [8]. Metabolomics data will be deposited in the MetaboLights repository of the EBI-EMBL [9]. Tool such as Registry of Open Access Repositories (ROAR) [10] and the Directory of Open Access Repositories (OpenDOAR) [11] will also be used when appropriate to increase access. Prior to repository deposit, MELCAYA data will be stored at the centralised database hosted by FCRB/HCB in a fully pseudoanonymised format with access strictly restricted to the investigator team.

3.3 Interoperability

Metadata will be based on controlled vocabularies, open data formats (CSV, XML, etc.) and secure data sharing protocols to make data easily retrieved, processed and re-usable by other systems.

3.4 Reusability

When and where appropriate, the project will distribute data through a Creative Commons Attribution International Public Licence (CC BY) or similar to enable reusability of the data while following the principle of “as open as possible, as closed as necessary".

4 Allocation of resources

The datasets will be deposited in the centralized repository hosted by FCRB-HCB and the different European open access research data repositories such as Zenodo, ISIC or EGA for at least 10 years after the conclusion of the project to guarantee continued accessibility and data discovery. No additional cost is expected for maintaining this data in the aforementioned repositories, but any unforeseen expenditures related to open access to research data will be covered by the EU funding under the conditions defined in MELCAYA’s Grant Agreement (GA). Data management related work will be supervised by the Project Coordination (PC) team and supported by the Ethics Monitoring Board (EMB) for ethical-legal considerations and the Exploitation Committee (EC) for intellectual property (IP) issues. Each MELCAYA partner will have the responsibility for the implementation of data management policies set out in this DMP. Dataset validation and registration of metadata and backing up of data shared through repositories (institutional-based or Zenodo) will be under the responsibility of the partner that generates the data.

5 Data security measures

In order to safeguard the rights of the data subjects, project partners will implement the following technical and organizational measures to ensure a level of security appropriate to the risk and sensitivity of the data transferred within the project, including:

- Following the pseudonymization procedures detailed in section 6.2.
- Keeping pseudoanonymised data and pseudonyms of patients separated.
- Encryption of data transferred if deemed necessary by local researchers.
- Limit the use of USB flash drives with a particular commitment not to store any personal data in those.
- Means to restore the availability of the data transferred and access to it within appropriate timeframes in the event of an incident.

- Physical and logical security (IT and communication networks) measures to protect the data transferred against accidental or illicit destruction, loss, alteration, disclosure or unauthorized access, including hacking or attempted hacking.
- Mechanisms to restrict and control access to data transferred, allowing individuals to be assigned access rights that are strictly necessary for their purposes.
- Keeping appropriate documentation on the processing activities.
- Obtaining the necessary certifications (particularly in terms of hosting health data if required by regulations).

Security measures regarding data archived in open access repositories such a Zenodo, OpenAIRE or other specialized European data repositories such as ISIC or EGA will be under the responsibility of the corresponding infrastructure owner.

6 Ethical and legal aspects

Most of the information detailed in the following sections has been already detailed in the ethics review, the ethics section of the Description of Action (DoA) and ethics deliverables of WP11. However, the most relevant information has been adapted to this document for easy referral.

6.1 Legal framework for data collection and processing

Two different approaches will be followed to deal with the processing of personal data within the framework of MELCAYA project. On the one hand, for **retrospective clinical studies,** the consortium has a lawful basis for the re-use of health data for scientific purposes under specified conditions and with adequate safeguards i.e., legitimate interests (article 6.1 (f) GDPR), combined with 'scientific research' article 9.2 (j) GDPR. In some cases, the data is obtained from registries in a completely anonymized format, which means that the subjects cannot be re-identified in any way and are therefore outside the scope of data protection law.

In the cases that the subjects could be re-identified, the guidelines on registry-based studies (EMA/426390) will be followed to ensure that access and use of the proposed data poses minimum to no risk to the study subjects or their fundamental rights and freedoms. If applicable, an authorization from the local Ethics Committee will be obtained. A protocol for the management of incidental or secondary findings detected during data processing and/or analysis will also be included (details in section 6.3) to ensure proper and secure handling of such information. In the cases where pre-existing ethics approvals are currently not in place, an authorization (or an amendment in the case of existing

approval) to access and use this data will be requested from each partner's respective local ethics committee or national competent body prior to study start-up.

On the other hand, for the **prospective studies,** in which new data will be collected from the data subjects during their clinical practice for research purposes, an explicit consent (article 9.2 (a) GPDR) will be required. The informed consent will explain to the subjects the purpose of the study as well as the nature and extent of their participation, complying with the rules set out in Regulation EU No. 536/2014. In the case of minors, consent will be obtained from the parent, guardian or legal representative. Subjects will be informed that participation in MELCAYA is completely voluntary, and that they can withdraw the consent at any time.

Finally, it is important to highlight that in all cases, when sharing collected personal data between MELCAYA partners, the parties involved will sign a **Joint Controller Agreement (JCA)** to ensure the protection of personal information. Conditions for the transfer of biological material (FFPE blocks, etc.) will be regulated by a **Material Transfer Agreement (MTA)** included in the Consortium Agreement (CA) complemented with correspondent traceability sheets for each sample exchange between partners.

6.2 Data pseudoanonymization

As previously commented, before uploading patient data to the eCRF (REDCap) hosted by the project coordinator, pseudoanonymization procedures will be implemented at each local data source center. Pseudonyms are character strings of defined length that are used instead of personally identifying data for linking different types of data of a study participant. Pseudonymization of data is an essential step in maintaining a high level of data protection, as it prevents re-identification of patients. In some cases, the data may already be stored under a pseudonym without the patients' identifying data (IDAT) being available. If available, patients' identifying data are documented at the local source centers. Each study participant is assigned a randomly selected pseudonym. Local patient lists are additionally maintained, which store the assignment of IDAT or local identifiers to pseudonyms. Both the local patient lists and the patient consents forms are stored locally and separately from the medical data at each source center. Without knowledge of the respective assignment of pseudonym and patient, no re-identification of individual persons is possible.

The development and evaluation of image processing AI algorithms, analyses of DNA sequences and methylation signatures or RNA expression profiles will be carried out exclusively by persons who had no direct patient contact during the data collection. Furthermore, the IDAT of the study participants will not be used for any purpose, but only the pseudonymized data of the study database. In the case of histopathological images, to decrease the possibility of tracing them back to their clinical origin via AI-based origin classification, developers at the German Center Cancer Research Center (DKFZ) will

apply an AI-algorithm for stain-to-stain translation that has been shown to significantly reduce the possibility to trace a histopathological image back to its original clinic [12].

Finally, for the protection of genomic data, MELCAYA will adhere to the Framework for Responsible Sharing of Genomic and Health-Related Data by the Global Alliance for Genomics and Health [13]. In order to allow centralized management of samples for genomics-level analyses on tissues or extracts, each sample's metadata associated with its local pseudonymous identifier will be entered in a biobank management system, with secured access controlled by an authentication system. Use of a laboratory information management system will provide project-level traceability of all samples included as well as of their transfers from one consortium member to another during the course of the project. Large -omics data sets will be made available to share within the consortium through the use of a project-specific pseudonymization standard.

6.3 Communication of incidental/secondary findings

Due to the potential for making an incidental/secondary finding in the framework of this project, the incorporation of a clause in the informed consent forms to let patients decide if they agree to be informed or not about these finding is absolutely necessary. The consent templates that are going to be prepared for the clinical studies will meet the following general requirements:

- Be written in the local language where the study is going to be performed and in terms that the patient can fully understand.
- Describe the aims, methods and implications of the research, nature of participation and any potential benefits/risks, etc.
- Explicitly state that participation is voluntary and every participant has the right to refuse and withdraw their participation, samples or data at any time without consequences.
- State how biological samples and data are going to be collected, re-used, destroyed, etc and for how long they will be stored.
- State what procedures will be implemented in the case of incidental findings, including an explicit recommendation for genetic counselling to prepare patients for follow-up testing and disclosure of a restricted list of potential molecular findings [14].

The clause added in the consent forms will clearly address the following aspects:

- That incidental or secondary findings may be discovered during the research studies.
- Whether the participant chooses to be notified or not about such findings.

- The process for communicating these findings, when the patient or their legal representative has opted in.

In the case of a potential incidental/secondary finding, the researcher is expected to inform an officer from his or her local Ethics Committee and coordinate a consultation with the medical professionals involved in the study from their participating institution to review and evaluate if the finding is relevant and how it should be communicated to the participant. In case of doubt, consultation can be made with other medical experts within the consortium. Contact with the patient would be done through the practitioner that generally attends the patient, using the available data recorded in the clinical history (if any). For minors, the general practitioner would contact with the parents or legal representatives (signatory of the informed consent). Ideally, a medical appointment would be scheduled when sharing this information to reassure the patient and avoid unnecessary stress.

The general conditions that must be always met to communicate an incidental/secondary finding are the following:

- It may affect a participant's health and welfare.
- It is scientifically and clinically valid.
- Ethical approvals have been obtained and the participant or their legal representative has opted in to receiving such results through their clinician(s) in the informed consent form.

Incidental and secondary findings will not be communicated:

- When the clinical information is anonymized, as it will be justifiably impractical or impossible to contact the research participant.
- When the participant has indicated that he/she does not want to be informed about such findings.

7 Commonalities within the *Understanding cancer* cluster

The European Union has put forward the *EU Missions* as a novelty of the *Horizon Europe* research and innovation programme for the years 2021-2027. The *EU Cancer Mission* has the goal (in combination with *Europe's Beating Cancer Plan*) of improving the lives of more than 2 million people by 2030 through prevention, cure and, for those affected by cancer (including their families), to live longer and better. The *Cancer Mission* board estimates a reduction in the expected mortality rates between 2021 and 2030 with respect to the baseline scenario (resulting from the current efforts of Member States) from 14 % to 20 % for females and from 30 % to 40 % for males.

The specific objectives of the mission are:

1. Understanding of cancer.
2. Prevention and early detection.
3. Diagnosis and treatment.
4. Quality of life for patients and their families.

MELCAYA belongs to the *Understanding* cluster (objective 1) financed by the *Horizon Europe* programme (HORIZON-MISS-2021-CANCER-02-03), which is aimed at better understanding the impact of risk factors and health determinants on the development and progression of cancer. The motivation for aiming at this objective is that, despite the huge advancements in the field of understanding cancer, much more research is still needed to realize why certain people, gender and age groups are at a higher risk of developing cancer, suffering from side-effects, etc. All these uncertainties limit the design of effective cancer prevention programmes as well as healthcare solutions adapted to each patient. Moreover, cancer research, healthcare providers, patient communities and industries are fragmented in the EU and do not benefit from patient engagement. The main points for collaboration identified so far between the projects within the cluster are the following:

7.1 Data re-use and generation & relation to the project's objectives

All the projects within the *Understanding* cancer cluster work on the integration of retrospective information from European registries, biobanks and cohort studies on different types of cancer combined with data from prospective cohorts. The processed information ranges from medical (clinical, epidemiological, histopathological, etc.) to environmental data (demographics, lifestyle, exposure to chemicals, etc.). Specifically, the projects will process:

- Clinical data (age, sex, medical history, tumour stage, …). The projects will re-use existing data mainly from clinical partners of the projects. The projects will also generate this kind of data in the prospective cohorts.
- Exposomics, including environmental, sociodemographic and lifestyle data (air pollution, chemicals, climate, diet, socioeconomic status, stress…). The projects will re-use existing data from various (mainly public) databases.
- Genomic data (measured genotypes, sequence data, gene expression, DNA methylation…). The projects will re-use existing data from different registries and biobanks as well as clinical partner databases. In most projects, this kind of data will also be generated in the prospective cohorts.
- Medical images (CT, MRI, PET, tissue slide image, histopathological and dermatoscopic images…). The projects will re-use existing data from different registries, biobanks and clinical partner databases. Some new imaging data will also be generated in some prospective cohorts.
- Volatile organic compound data. Several of the projects will generate data using sensor devices (sensing patch, breath analyser, spectrometry on card).

The data re-use and generation in the projects is intrinsically linked with the projects' objectives. All projects in the cluster have the common objective of understanding the cancers they research, obtaining insights into the risk factors, causal pathways and the mechanism of development of the cancers, improving diagnosis and prognosis. Specifically, the projects have the following aims for re-using and generating the above-mentioned data in common:

- Identification of environmental, lifestyle and genetic risk factors for the different cancers, and later, biological mechanisms through which the identified risk factors promote carcinogenesis. For this analysis, clinical data will be combined with exposomics, and often in a later stage, with genomic data, to obtain polygenic risk scores.
- Development of imaging models and AI-based diagnostic tools.
- Implementation of rapid, cost-effective and non-invasive sensor tools for early detection.

More specific information about the commonalities in the data re-used and generated in the different projects will be clarified as the projects progress, and will be provided in later versions of the DMPs.

7.2 FAIR data management

The data used and generated can be useful beyond the scope of the projects, notably to healthcare professionals and cancer researchers wanting to understand risk factors for the cancers and diagnose

cancer at earlier stages. Indeed, the data generated within each project can be useful to the other projects of the *Understanding* cluster too.

The overarching goal of this common chapter is to find common practices to share the information in pan-European research infrastructures, such as the European biobanking platform (BBMRI-ERIC) or the future UNCAN.eu platform, a federated cancer data hub platform currently under development. This is a particularly critical point, as at the present time, patient health data networks in Europe show a high level of heterogeneity in terms of involvement of EU Member States as well as the types and interoperability of collected data, organisation and governance of data storage, security or the possibility to use this data for research purposes.

Therefore, all of the projects in the cluster are committed to manage their data according to the FAIR principles. The projects are planning on implementing the following measures to make the data findable, accessible, interoperable and re-usable:

- To ensure and optimise findability of the data, all data will be assigned a unique and persistent identifier.
- Defining metadata will be particularly important throughout the projects, using controlled vocabularies, non-proprietary / open data formats and secure protocols for data transfer and sharing. Metadata will be made available clearly referring to the identifier of the described data.
- Where possible, the data used and generated in the projects will be deposited in open repositories. The projects are considering using generalist data repositories within the Open Access Infrastructure for Research in Europe, such as OpenAIRE or Zenodo, in order to ensure the widest degree of public access possible. More specific types of data will be deposited in suitable discipline-specific repositories.
- The projects will also use open community data standards and models where possible. For example, the use of Observational Medical Outcomes Partnership (OMOP) as data model and Health Level Seven as standard for exchanging clinical data are being considered for health data interoperability.

Moreover, the projects are considering the possibilities for common exploitation of data within the cluster. We are examining the following possibilities within the cluster:

- Sharing data during the projects.
- Sharing risk scores and models near the end of the projects.
- Publishing a common paper.

- Implementing the results in healthcare policies and screening programmes.

A data management board in the *Understanding* cluster has been established to address commonalities on data standards, data validation, as well as on the best practices regarding data privacy (pseudonymization or anonymization techniques), storage and exchange protocols. At this early stage, our group has met regularly and focused on the basic commonalities mentioned above. Our collaboration will continue with regular meetings and email communications. Our group will also meet during the annual cluster meeting, closely organized with the European Commission to address common scientific challenges. During our collaboration, we will further explore the intentions mentioned above and will provide more detailed versions of the common aspects of data management in the *Understanding* cluster in upcoming versions of the DMPs.

8 Conclusions

In this report, we have reviewed the main aspects related to the data management activities that will be required throughout MELCAYA project. First, we have identified the main datasets to be managed during the project lifecycle, including aspects such as data type, format or storage, the data management workflow, what is the purpose of data collection and how it is strictly linked to the aims of MELCAYA. In the next section, the strategies to make data findable, accessible, interoperable and reusable (FAIR) have been presented, as well as an explanation on resource allocation to ensure the data management approach can be followed not only during but also after project completion. The general technical and organizational measures to ensure an adequate level of data security are also presented, as well as an overview of the main ethical and legal aspects regarding data collection processing, including legal framework, data pseudoanonymization techniques and how incidental and secondary findings are going to be communicated. Finally, a brief description of the common areas for collaboration in terms of data management with other projects of the *Understanding cancer* cluster is provided.

References

[1] https://projectredcap.org/software/

[2] https://www.nature.com/articles/sdata201618

[3] http://www.openaire.eu/

[4] https://www.zenodo.org/

[5] https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

[6] https://digital-strategy.ec.europa.eu/en/policies/cancer-imaging

[7] https://ega-archive.org/

[8] https://www.ncbi.nlm.nih.gov/geo/

[9] https://www.ebi.ac.uk/metabolights/

[10] http://roar.eprints.org/

[11] https://v2.sherpa.ac.uk/opendoar/

[12] Brinker et al., *Multi-domain stain normalization for digital pathology: a cycle-consistent adversarial network for whole slide images*, arXiv preprint (2023)

[13] Framework for Responsible Sharing of Genomic and Health-Related Data: https://www.ga4gh.org/genomic-data-toolkit/regulatory-ethics-toolkit/framework-for-responsible-sharing-of-genomic-and-health-related-data/

[14] Miller et al., Recommendations for reporting of secondary findings in clinical exome and genome sequencing, Genetics in Medicine (2017)