Call: HORIZON-HLTH-2022-TOOL-11
Topic: HORIZON-HLTH-2022-TOOL-11-02
Funding Scheme: HORIZON Research and Innovation Actions (RIA)

Grant Agreement no: 101095435

Deliverable No.143

D4.2 Data Management Plan V1.0

Contractual Submission Date: 30/June/2023

Actual Submission Date: 30/June/2023

Responsible partner: P1: Maastricht University (Chang Sun)

Funded by
the European Union

Grant agreement no.	101095435
Project full title	REALM - Real-world-data Enabled Assessment for heaLth regulatory decision-Making

Deliverable number	**D4.2**
Deliverable title	**Data Management Plan V1.0**
Type[1]	R
Dissemination level[2]	PU
Work package number	WP-4
Work package leader	Chang Sun (P1-Maastricht University)
Author(s)	Chang SunMichel Dumontier, Birgit Wouters (P1-UM), Ilias Siniosoglou P3-MINDS, Bart Elen (P6-VITO).
Keywords	Data Management Plan, FAIR, Open Data, Security, Confidentiality, Governance, Compliance

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 European Health and Digital Executive Agency (HaDEA).

Neither the European Union nor the granting authority can be held responsible for them.

Document History

Version	Date	Description
V1.0	30/June/2023	First submission of Data Management Plan (DMP) for REALM

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)
PU: Public, fully open, e.g. web
SEN: Sensitive, limited under conditions of the Grant Agreement

Table of Contents

List of Abbreviations and definitions

REALM - Real-world data Enabled Assessment for heaLth regulatory decision-Making

DMP - Data Management Plan

EHR - Electronic Health Record

SUS - System usability scale

SOAP note - Subjective, Objective, Assessment and Plan (SOAP) note

CT Scan - Computerized Tomography scan

DICOM - Digital Imaging and Communications in Medicine

GDPR - EU General Data Protection Regulation

FHIR - Fast Healthcare Interoperability Resources

FAIR - Finable, Accessible, Interoperable, Reusable

OMOP - Observational Medical Outcomes Partnership

Executive Summary

This data management plan (DMP) outlines the strategies, processes, and resources employed to ensure effective management of data throughout the lifecycle of REALM project. The key objectives of the DMP are to promote data integrity, facilitate data sharing and reuse, and ensure compliance with relevant regulations in Europe and funding agency requirements from Horizon Europe. To provide a comprehensive data management overview, this DMP took the inputs from five demonstrators who will provide real-world data and use cases in the project and considered the feedback from members working in WP1 (Michel Dumontier), WP2 (Birgit Wouters), WP3 (Ilias Siniosoglou), and WP5 (Bart Elen).

The goal of REALM project is to create an inclusive platform for evaluation and certification of software in healthcare where both the developers and the regulatory bodies have access to a standardized set of technology stack and data. The research activities include building REALM architecture, establishing real-world and synthetic data repository, integrating REALM tools and testing and monitoring REALM tools. In REALM, data plays a crucial role in achieving the project objectives, robust data management practices have been implemented to maximize the value and impact of the generated research outputs.

This DMP describes the types and formats of data, the size of the data, origin and provenance of the data that will be re-used and generated in the project. Then, the DMP elaborates how the data and other research outcomes in REALM project will be made Findable, Accessible, Interoperable, and Reusable. We classify data into 1) real-world data (that are individually identifiable) from five demonstrators in the project, 2) anonymize data (that are not individually identifiable) which are requested from biobanks and data repositories (such as UK Biobank, FINNGEN), and 3) synthetic data (that are not individually identifiable) which are generated based on the real-world data. Based on the different nature of the data, the project proposes different strategies to make these data accessible and reusable for other parties under different level of restrictions.

Data storage and security are given utmost importance in the project. The data is stored in private clouds or trusted data repositories which provides secure and reliable storage solutions. Robust data security measures, including encryption, access controls, and regular backups, will be implemented to protect the integrity and confidentiality of the data. Long-term preservation of the data is crucial for future reference and potential reuse. The project uses persistent identifiers for all datasets and research outcomes and archives and stores data using the trusted services such as university or national research infrastructure to ensure the longevity and accessibility of the data beyond the project duration. The DMP adheres to relevant compliance and ethical considerations. This involves EU GDPR, and national data protection regulations in state members, informed consent procedures, or anonymization/de-identification methods to protect the privacy and confidentiality of individuals or entities involved in the research.

In conclusion, this first version of DMP for REALM project provides a comprehensive overview of managing data in the project. The strategies and practices outlined in the DMP ensure data integrity, facilitate data sharing and reuse, and adhere to relevant compliance and ethical considerations. The achievements attained through the implementation of the DMP highlight the project's commitment to robust data management practices and its contribution to advancing scientific knowledge.

After this first version of DMP (V1.0), we will keep updating the DMP every six months throughout the project lifecycle. The new data, methods, software, and other research outputs that are created from all the work packages in REALM must be added in the updated version of DMP (V2.0). Any major changes, deletion, or extension from the first version of DMP will be highlighted in the updated DMP (V2.0).

1 Introduction and Data Summary

The Horizon Europe Model Grant Agreement requires that a Data Management Plan ('DMP') is established and regularly updated. This DMP is based on the template[3] that is recommended for Horizon Europe beneficiaries. By completing the sections of the template, the requirements for research data management of Horizon Europe as described in article 17 and analysed in the Annotated Grant Agreement, are all addressed.

1.1 Re-use of existing data

In the overall project, we will re-use and collect existing demographic data, electronic health records (EHR), medical imaging data (e.g., CT Scan), medical text data (e.g., SOAP note, a communication document between health professionals), medical sensor data (e.g., heart monitoring data), genomic data, and questionnaire data. REALM will start with requesting the access to publicly available data from various national EU and UK data repositories including but not limited to Estonian Biobank (https://genomics.ut.ee/en/content/estonian-biobank), FINNGEN (https://www.finngen.fi/en), UK Biobank (https://www.ukbiobank.ac.uk/), Lifelines (https://www.lifelines.nl/), Health RI (https://www.health-ri.nl/). Then, we will set up bilateral agreements and request access to use anonymized patients' data from hospitals, clinics, research institutes, and other healthcare providers in the Netherlands, Belgium, and Germany (refer to Figure 4: REALM Component Architecture in the REALM Grant Agreement). To evaluate the REALM infrastructure, the project includes five demonstrators as use cases using different types of data. The five demonstrators will re-use some existing data collected from their studies and devices and will also collect and generate new data in the project.

1. **DuneAI** - from partner Maastricht University – is an automated segmentation software based on deep learning techniques that performs state-of-the- art detection and volumetric (3D) segmentation of non-small-cell lung cancer (NSCLC) tumours in CT scans. DuneAI will re-use existing patients' CT Scan data and collect new patients' demographics, CT scans, System Usability Scale (SUS) evaluations data, and qualitative feedback from clinicians.
2. **PGX2P** – from partner: VITO - tests a set of germline genetic variants representing a new model for employing stratified medicine in practice. PGX2P will re-use pharmacogenomics data from UK Biobank for development purpose and collect new real-world data for evaluation purpose within REALM infrastructure.
3. **STAR** – from partner: Liege University – is a model-based decision-support system to predict blood glucose ranges for insulin and nutrition intervention. STAR will collect new clinical patients' data including their demographic and clinical data focusing on glucose and patient response to insulin and questionnaire data for usability assessment.
4. **ASCOPD** - from partner: TRAQBEAT – uses Traqbeat sensor and methods to measure biomarkers. ASCOPD will re-use the existing patients' data including demographic and clinical data from their admission and discharge events and generate new heart-related data using TRAQBEAT sensors.
5. **COPowereD** - from partner: COMUNICARE – is a generic app to follow up with COPD patients and predict the detection of COPD exacerbations. COPowereD will collect new observation data from patients with chronic obstructive pulmonary diseases from hospitals.

REALM will collect and use sensitive and individually identifiable datasets which cannot be accessed and shared by third parties or can only be accessed and shared with restrictions. In case of restricted data access, we will create and optimize synthetic data to be realistic at the individual level and representative at the population level based on the real-world data. The generated synthetic will be used for testing and experimenting the

[3] HE DMP Template: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

development of REALM infrastructure ensure the REALM architecture can be built without a delay in the project timeframe. At the later stage of the project, the REALM infrastructure will need to be evaluated and assessed by the regulatory authorities, certification authorities and medical device manufacturers, and other relevant stakeholders. Questionnaire data about the use of the REALM infrastructure will be generated and collected for the optimization purpose.

1.2 Types and formats of data

Table 1. An overview of data types, formats, and elements in REALM

Data Types	Formats	More information about data
Demographic	Structured data	Age, sex, ethnicity, address, socio-economic status, education
Electronic Health data (EHR)	Structured data	Diagnosis, medication data, diabetes status, hba1c, mortality, ICU length of stay, smoking, allergies, vaccination status, BMI, insulin rates, nutrition rates, glycaemia, nutrition target, vital signs, symptoms, medical history, lab tests on admission and discharge, and other clinical information.
Imaging data	DICOM, PNG, JPG	CT Scans (medical images of patients' lungs for tumour segmentation), Chest X-rays (medical images of patients' hearts)
Sensor data	Time-series data	Heart and lung monitoring data: Vital signs (such as ECG, heart rate, heart rate variability, blood pressure, respiratory rate, SpO2, skin temperature
Genomics data	Sequencing data	Genotyping, exome sequencing, whole genome sequencing, etc
System Usability Scale evaluations	Sequencing data	Numerical scores and potentially written comments from the system usability scale assessments.
Qualitative Feedback	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews

Table 2. Descriptions of datasets from five demonstrators in REALM

Demonstrator	Data Types	Formats	Size/samples	Elements
Dune AI	Demographics & EHR	Structured	Based on 1328 samples	Age, sex, disease, and other clinical information
	CT Scans	Imaging	600GB (1328 * 500 MB)	Medical images of patients' lungs used by the DuneAI software for tumour segmentation in DICOM format.
	SUS evaluations	Structured	-	Numerical scores and potentially written comments from the System Usability Scale assessments.
	Qualitative Feedback	Text	-	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews
PGX2P	Demographics & EHR	Structured	Based on 100 samples	Sex, age, ethnicity, diagnosis, and medication data
	Genomics	Sequencing	20-50TB	Genotyping, Exome Sequencing, Whole Genome Sequencing
	Genotyping	Sequencing	100 GB (100 samples* 1 GB)	Collect from a small cohort for validating the application.
STAR	Demographics & EHR	Structured	60MB (20 samples * 3 MB)	Age, sex, BMI, diabetes status, hba1c, severity score, mortality, ICU length of stay, insulin rates, nutrition rates, glycaemia, nutrition target, and potential other additional information affecting glucose control
ASCOPD	Demographics & EHR	Structured	Based on 200-300 samples	Date of birth, address, smoking, allergies, vaccination status, BMI, medical history, Lab tests on admission discharge, physical activity, smoking status, weight.
	Heart monitoring	Time-series	-	Vital signs: ECG, heart rate, heart rate variability, blood pressure, respiratory rate, SpO2, skin temperature
	Chest X-Ray	Imaging	-	X-Ray images
COPowereD	Demographics & EHR	Structured	-	Age, sex, disease, symptoms, and other clinical information
	Lung monitoring	Time-series	100 MB	Vital signs

1.3 Purpose of data generation and re-use

REALM project is creating an inclusive platform for evaluation and certification of software in healthcare where both the developers and the regulatory bodies have access to a standardized set of technology stack and data. We will re-use the existing data from data repositories and sources with public or on-request access to develop and test the platform. Then, we will generate synthetic data using real-world health data from the demonstrators in the projects to evaluate and optimize the REALM platform. Finally, to validate REALM platform in practice, we will apply the real-world health data to the REALM platform and collaborate regulatory authorities, certification authorities and medical device manufacturers to assess REALM platform and collect feedback and suggestions to and make improvements.

1.4 Size of data that will be generated or re-use

The expected size of the real-world raw data in the overall project will be about 20 – 50 TB which is mainly from the genomics data and medical imaging data. The demographic data, EHR data, and monitoring data (vital signs) are expected to be 5-10 GB in total. Regarding the number of data subjects (e.g., patients, clinicians, physicians), the project expects to re-use data from BioBanks and repositories from around 200,000 data subjects (around 100,000 data subjects from UK Biobank, 50,000 from clinical databases such as MIMIC III, and round 50,000 from other biobanks and repositories).

1.5 Origin/provenance of the data

In the developing and training stage of REALM infrastructure, the origins/provenances of data are the data sources such as UK BioBank, data repository owners, and hospitals and research institutes that collect the original data. In the evaluation phase, the origin/provenance of data in five demonstrators are also from the data sources where the patients are treated such as hospitals in the Netherlands (Dune AI), hospital and research institutes and hospitals in Belgium (PGX2P, STAR, COPowereD), and hospitals in Greece (ASCOPD).

1.6 Usefulness of data for other parties

At the beginning stage of the project, we foresee the data might be useful for researchers and developers who are working on evaluation and assessment of healthcare devices and software, and as well as regulatory authorities, certification authorities, and medical device manufacturers. In the REALM, Task 3.1 (Design and Governance of the Federated REALM Architecture) and Task 4.2 (Standard Data Quality Framework) will identify the stakeholders and actors for REALM and analyse their user requirements, security and privacy requirements. Therefore, we expect a more specific and clearer overview of the data usefulness for other parties in the next (updated) version of DMP.

2 FAIR Data

2.1 Making data findable, including provisions for metadata

2.1.1 Persistent Identifier of data

Persistent identifiers are those that can withstand social and technological challenges including a change in the responsible party or a change in the (HTTP) location. Our strategy will be to rely on persistent identifiers assigned and maintained by 3rd parties, namely trusted repositories or data service providers cooperating with European Persistent Identifier Consortium. More specifically, we will plan to make use of Figshare (https://figshare.org) and PhysioNet (https://physionet.org), which both offer Digital Object Identifiers for submitted data. Where necessary, we will also consider other forms of persistent identifers such as the W3C Identifier (https://w3id.org) to redirect from a persistent identifier to another location (e.g., GitHub Release package, web API, etc).

2.1.2 Metadata of data

The datasets will be attached with rich metadata that is machine- and human- readable so that the datasets can be searchable in the common search engines (such as Google, Bing). We will apply the existing metadata standards to create metadata for each type of data we used in the project. We consider HCLS Dataset Description Guideline (https://www.w3.org/TR/hcls-dataset/) Dublin Core Metadata standard (https://www.dublincore.org/specifications/dublin-core/dcmi-terms/), DCAT – Data Catalog Vocabulary (http://www.w3.org/TR/vocab-dcat/), Schema.org (https://schema.org/), and RO-Crate (https://www.researchobject.org/ro-crate/) for describing project data. Since the datasets in the project are mostly health, biology, and life science related, we also consider to apply metadata standards including (not limited) Access to Biological Collection Data Schema (ABCD, https://abcd.tdwg.org/), Darwin Core (http://rs.tdwg.org/dwc/index.htm), Genome Metadata (http://enews.patricbrc.org/faqs/genome-metadata-faqs/), and Bioschemas (https://bioschemas.org/).

2.1.3 Keywords in metadata

The relevant and searchable keywords will be provided in the metadata. The metadata will be structured using the standards and technologies that can optimize the searching possibility in research repositories and in Google, Bing, and other common search engines.

2.1.4 Harvestable and indexable metadata

For the datasets that will be uploaded to open-access data repository (such as Figshare), the projects will provide comprehensive metadata based on the requirements of the data repository. Using the research data repositories, we expect these metadata can be harvest and indexed. In any case, if some sensitive datasets are not uploaded to a data repository but maintained in a controlled environment, we will provide all necessary metadata in a human- and machine-readable manner so that the metadata can be harvested and indexed.

2.2 Making data accessible

2.2.1 Trusted data repository

The choice of trusted data repository will be based on the sensitivity of the data used in the project. The project will deposit the generated synthetic data that is evaluated as “non-personal data” in an appropriate repository (see Section 5.2 on choice of data repositories). The real-world data will be deposited in a trusted repository after being anonymized and generalized to ensure the processed data is not individually identifiable. The sensitive real-world data and synthetic data that are still considered as “personal data” will be managed and maintained at their source site in controlled environments. The metadata of these datasets will be published in a trusted repository as well as the requirements and procedure of requesting the access to the data.

2.2.2 Exploring arrangement with the identified repository

For the regular non-sensitive and non-identifiable real-world data and synthetic data, the project intends to use open-access repository such as Figshare and Physionet to deposit. Data that are not individually identifiable but contains sensitive information such as pseudonymized clinical data, the project intends to deposit them in trusted repository using services like SURF. For sensitive and individually identifiable data such as CT Scans and genomic data, the project intends to keep them in the secure storage at their sources. However, the metadata of these data should be published and stored in the trusted repository.

2.2.3 Data identifiers in the repository

The project will only use the data repositories and services that can provide the persistent identifiers and assign a unique persistent identifier to a digital object.

2.2.4 Data availability

REALM project will include health research data from biobanks and data repositories, synthetic data generated from real-world patient data, and real-world patient data. REALM will keep the existing published research data

at their original sites (e.g., UK Biobank) and indicate the sources, request time, and procedures publicly. The anonymized real-world data and synthetic data that are not individually identifiable will be provided publicly available for research. The pseudonymized real-world data that are not individually identifiable but seen as sensitive data can only be made available for other parties with restriction. Where necessary, data sharing and data processing agreements, confidential agreements will be established by the data controllers based on EU data protection laws and GDPR.

2.2.5 Potential embargo

An embargo refers to a specified period during which access to certain data or information is restricted or delayed. During this embargo period, the data or information is not made available to the public or other researchers. The purpose of imposing an embargo is to provide time for activities such as publishing research findings or seeking intellectual property protection, such as filing a patent. An embargo period may provide the necessary time to identify improvements and to file for any necessary patents within REALM infrastructure and in the demonstrators' use cases. The exact embargo time will need to be determined in consultation with REALM stakeholders. The possible time range could be around 6-24 months.

2.2.6 Data accessibility through a free and standardized access protocol

The data or metadata in REALM will be accessible to anyone with a digital device, an internet connection, a web browser, and permission to access the data where restrictions exist. In using web accessible data repositories as our primary means of disseminating research products, we expect that the primary access protocol for both data and their metadata will be via the Hypertext Transfer Protocol (HTTP) and HTTP Secure (HTTPS). Where data access restrictions are in operation, we anticipate the use of standardized protocols for authentication (e.g., passwords, multi-factor authentication) and authorization (e.g., OAUTH2). Other free and standardized access protocols sometimes used for data archive interactions including the Interplanetary File System (IPFS), Amazon Simple Storage Service (S3), RSYNC, and the File Transfer Protocol (FTP). Where possible, we will seek solutions that provide multiple pathways to access our datasets.

2.2.7 Restricted data access

Conditions will be applied to restrict the access of individual-linked real-world data. We will document the procedures for each source data site and devise a protocol for REALM generated content. A preliminary workflow, subject to further discussion is: A data request application is submitted to the Data Access Committee indicating the purpose, the research team and their track record, the required data elements and their justification, the potential for impact, the timeframe for analysis; the application is then reviewed by some authorized individual/group; a decision on approval is made in a timely fashion and subsequently communicated to the applicant; positive approval then requires completion of a data use agreement ; and follow-up on completion of study/expiry of data use agreement.

2.2.8 Identity of data requester

In the restricted data use/sharing, the identity of the person accessing the data will be ascertained through an identification process involving a combination of username/password authentication, two-factor authentication, and institutional verification. The credentials of the person requesting access, such as their institutional affiliation and the purpose of their research, will also be verified before access is granted.

2.2.9 Data access committee

In the restricted data use/sharing, the project intends to have a data access committee that includes representatives from the REALM consortium, participating hospitals, and independent ethics experts. The primary role of the data access committee is to review and evaluate requests for data access, ensuring that the data is only used appropriately and in accordance with relevant laws and regulations (e.g., EU General Data Protection Law), policies, ethical considerations, and consortium's data use agreements, and the data is only

shared with parties that have a legitimate research or development need. The committee should also ensure that appropriate measures are in place to protect patient privacy and the proprietary aspect.

2.2.10 Metadata availability

The metadata of all the datasets will be made openly available and licenced CC0 Public Domain Dedication or CC-BY 4.0 license (note that the licences are applied to the metadata rather than data). We will clearly and detailed describe the availability of the data and how to request the access to the data.

2.2.11 Availability period

The data will remain available and findable for ten years and metadata will remain available even after data is no longer available.

2.2.12 Software requirements for accessing data

The structured data will be stored in a regular relational database and can be exported to CSV files. The medical imaging data and sequence data can be loaded using free and open programming language such as Python. The instructions of how to access and read the data will be provided clearly and publicly.

2.3 Making data interoperable

2.3.1 Data and metadata vocabularies

All re-used and generated data in the REALM project will largely adhere to commonly accepted standards within the medical and research communities to ensure interoperability and reusability. The demographics and clinical data will be structured according to the standardized formats for Electronic Health Records (EHR) such as HL7 standards[4], Fast Healthcare Interoperability Resources (FHIR)[5], Observational Health Data Sciences and Informatics (OMOP)[6], and other standard healthcare/medicine specification. The medical imaging data such as CT scans will be structured in DICOM (Digital Imaging and Communications in Medicine) or PNG or JPG imaging format for processing. The genomic data will follow the established standards in GA4GH (Global Alliance for Genomics and Health) or Pharmacogenomics Research Network (PGRN). For the questionnaire or survey data for infrastructure evaluation, there are no specific standards, but the project will use a standard format for the representation of the survey results, like CSV or Excel, accompanied by metadata explaining the context and meaning of the data. Finally, for the datasets that has no well-known standards (to our best knowledge), the project will take the recommendations from the Refined eHealth European Interoperability Framework (EIF) which has a set of standards, protocols, procedures, and policies for the interoperability of eHealth applications within the EU. At this stage of the project, we do not anticipate using any uncommon or generate project specific ontologies or vocabularies.

2.3.2 Linking data with other resources

The project will re-use existing datasets from data repositories and may link with newly generated or collected data from the project. The project will ensure the data can be linked using common/shared ontologies and vocabularies.

2.4 Increase data re-use

2.4.1 Data documentation for re-use

We will provide necessary documentation about the data analysis models by adding a README file and WIKI page in the code repository including the vocabulary of the variables, collection methods, detailed description of the pre-processing of the data, and analysis methodology.

[4] *HL7 standards: https://www.hl7.org/implement/standards/*
[5] *FHIR: https://www.hl7.org/fhir/*
[6] *OMOP: https://www.ohdsi.org/*

2.4.2 Data availability for widest re-use

Only the non-sensitive synthetic data may be freely available in the public domain with Commons Attribution - Non-Commercial Use - No Derivatives (CC BY-NC-ND) license. The real-world data used in REALM includes sensitive health and genetic information and personal identifiable information. Therefore, given the nature of the data and national laws governing personal data, it's not appropriate to make these data available without access restrictions.

2.4.3 Data re-usability during and after the project

Synthetic data generated from REALM project which are individually unidentifiable can be useable by third parties during and after the project. The anonymized real-world data which are individually unidentifiable can be useable by third parties under certain conditions such as approval by the data access committee, signing data use agreement, and access through secure and controlled environment. The sensitive and individually identifiable real-world data will not be usable by third parties.

2.4.4 Data provenance

We will capture and store the provenance for all source and generated data. The provenance information will include the responsible individual/organization and their contact information, the dates in which the data were collected/generated/used, the tools/workflows to collect/generate the data. We will use common provenance standards such as PROV-O to capture basic provenance and look augment this with other provenance guidelines (e.g., HCLS Dataset Description). Data versioning will be supported, where possible, through a Git-style version control system and logging. The data provenance will be stored and archived for citable reproducibility. For synthetic data, we will document the metadata of the training data, the structure and hyperparameters of the generative models, the versions of the generated data, and all the relevant information of the generators and synthetic data using ML/Workflow metadata specifications such as FAIR Workflows (https://fair-workflows.github.io/project.html).

2.4.5 Data quality assurance

The project will first establish clear criteria and standards to evaluate the data quality based on different types of data (e.g., clinical data, medical imaging data, or genomic data). The criteria include accuracy, completeness, consistency, reliability, timeliness, and adherence to data collection protocols. Then, the project will conduct the periodic reviews or assessments on samples of data to ensure the data quality in the project. We will document the reviewed data, review results, identified data quality issues, and any changes or corrections to fix the issues.

3 Other research outputs

The research outputs in REALM project can be categorized into **1) infrastructure or models** where the design and code are the main outcome (such as D3.1 Federated REALM Architecture, D4.6 synthetic data generator and D5.1 REALM intelligent analytics dashboard, **2) guidelines or protocols** where the requirements, principles, and guidelines are the key outcome (such as D3.2 Guidelines on AI based MDSW evaluation, D4.4Standard Data Quality Framework), and 3) **scientific publications** such as dissertations, posters, oral presentation slides.

There are two sensitivity levels of the outputs that are indicated in the REALM grant agreement Pages 22-28 – public (PU) and sensitive (SEN). For the outputs that can be public, the code of the developed infrastructure and models will be open source, uploaded to a public repository (such as Github), and maintained during and after the project, while the report and documentations will be published on the project website. To communicate the project findings with research community and the public, we intend to publish scientific articles, posters, and presentation slides in journals, press, conferences. We will aim at publishing with open-access journals and/or through the Open Research Europe (ORE) platform[7]. A high-level strategy will be agreed upon by the partners

Open Research Europe Platform: https://open-research-europe.ec.europa.eu/

regarding the content of publications and other communication activities and this will be in line with the Consortium Agreement (CA), which deals with issues related to IP ownership and similar matters.

For the outputs that are regarded as sensitive, we will publish the methodology and pseudo code of the infrastructure and models in a repository, and we will describe the requirements and process of requesting the full code (if applicable). For the reports and documentation with the sensitive level, we will publish the high-level abstracts on project website and indicate the requirements and process of requesting the full text (if application).

4 Allocation of resources

4.1 Estimated costs for making data and research output FAIR

At this stage of the project, we estimate the data management related costs will be generated by personnel costs for making data and research output FAIR, data application fees by requesting data from biobanks and repositories, service fees for data storage, transfer, and security. We estimate the total amount of cost may be approximately 276,000 euros including personnel costs, service fees, and equipment costs.

4.2 Resources

We will allocate some personnel costs to WP4 (Real-world data and synthetic data repository), particularly T4.1 (The federated cloud-based data repository catalogue and management) and T4.2 (Standard Data Quality Framework) to manage data and make the data and research outputs FAIR in the project. The costs related to data requests, storage, transferring, and security during the project time will be covered by the budget under category of “Equipment” and “Other goods and services”. The costs of collecting and managing real-world data which are from the five demonstrators are covered by the demonstrators themselves.

4.3 Responsible parties for data management

In the development of REALM infrastructure, real-world data comes from the five demonstrators. These demonstrators are responsible for their own real-world data as the data controllers. They are also responsible for the synthetic data that are generated based on their datasets. The role of WP4 - Real-world and synthetic data repository is to coordinate data requirements from other work packages and gather information about data characteristics from the demonstrators. WP4 is also responsible for delivering generative methods but for the generated datasets. The data management of existing research data from biobanks and public data repositories within REALM is assigned to WP4.

4.4 Long term preservation

The metadata of all the data in REALM project will be preserved persistently at trusted repositories and university servers. The research data from biobanks or data repositories are maintained by their respective organizations. However, we will record the data request process and metadata of the re-used data. The synthetic data will be preserved for ten years for scientific research use. The real-world data will stay at the data sources and will be preserved by their own data controllers.

5 Data security

5.1 Provisions for data security

The non-sensitive synthetic data will be stored in trusted/certified research data repositories. The security standards of the repository will apply in this case. The real-world data in the project will be pseudonymized and stored securely in encrypted servers with access controlled by strict authentication mechanisms such as university servers or private cloud infrastructure. We will back up the data periodically to prevent data loss and ensure data recovery if needed. Sensitive and individually identifiable real-world data can stay at their sources in a secure encrypted storage. Then the security standards of the data sources will be applied in this case. The data transfer

may happen, if necessary, through encrypted channels. The minimum necessary amount of data can be transferred after anonymisation or pseudonymisation.

5.2 Data storage, preservation and curation

All the data including real-world data and synthetic data will be safely stored in trusted repositories for long-term preservation and curation with persistent identifiers and versioning. In choosing a repository, we consider the factors including the nature and the sizes of data. Due to the different nature of data used in the project, we consider storing non-sensitive synthetic data in open-access and public research repository such as Figshare. Sensitive and pseudonymized data (which are individually unidentifiable) can be stored in research data repositories such as Physionet or Dryad with restricted data access control. The sensitive and individually identifiable data which cannot be shared or can be shared only with data use agreement will be stored in secure private cloud or proprietary repositories in the universities or partner companies. For special data such as genomic data, the project considers using standard services such as European Genome-phenome Archive (EGA) to permanently archive and share personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects. For a large amount of dataset, we will make an agreement within project to pay for storing the data for a long term or persistently.

6 Ethics

6.1 Possible ethics or legal issues on data sharing

The real-world patient data in REALM including but not limited to patient demographics, clinical data, laboratory results and CT scans, genomics are categorized as sensitive personal identifiable data in the General Data Protection Regulation (GDPR). GDPR or new regulations may hamper the potential re-use of the data in the future. All the data are and will be collected with consent and will be de-identified and pseudonymized before sharing to mitigate the risk of privacy breach. The data that is from or for clinical trials will need an ethical approval from the review committee for data collection and sharing. The assessment questionnaire data from clinicals and physicians needs informed consent for their data. The evaluation and feedback data of proprietary software or devices using REALM infrastructure may potentially reveal information about the functionality of the software or devices. REALM Work Package 2 (Ethical, Legal & Societal Framework) is working on providing robust and socially responsible ethical-legal standards to address the challenges of developing and implementing REALM infrastructure and tool in clinical practice. Therefore, WP2 will systematically study and address the possible ethics and legal issues on data sharing in REALM.

6.2 Informed consent

All re-used, prospective data and retrospective data will be used and collected with proper informed consent from data subjects for data sharing and long-term preservation.

7 Conclusion

In conclusion, the DMP for the REALM project describes the strategies of re-use, collection, and management of various types of data, including demographic data, electronic health records (EHR), medical imaging data, medical text data, medical sensor data, genomic data, and questionnaire data. The project will request access to publicly available data from national EU and UK data repositories and establish bilateral agreements to access anonymized patients' data from hospitals, research institutes, and healthcare providers in the Netherlands, Belgium, and Germany. The DMP integrates the descriptions of real-world data from five demonstrators in REALM that will re-use existing data from their studies and devices, as well as collect and generate new data. These demonstrators cover different use cases and involve data such as CT scans, pharmacogenomics data, clinical data, heart-related data, and observation data from patients with chronic obstructive pulmonary diseases. To ensure the development and testing of the REALM infrastructure, synthetic data will be created and optimized to be

representative of real-world data at the individual and population levels. This DMP covers different strategies for data with different levels of identifiability - personally identifiable real-world data, anonymized research data, and synthetic data.

To ensure the FAIR (Findable, Accessible, Interoperable, and Reusable) principles of the data, the project will assign persistent identifiers to the datasets and attach rich metadata to make them findable. Metadata standards such as HCLS Dataset Description Guideline, Dublin Core Metadata, and DCAT - Data Catalog Vocabulary will be applied. Harvestable and indexable metadata will be provided, and keywords will be included to optimize searchability. Depending on the sensitivity of the data, trusted data repositories (such as Figshare or Physionet), secure and/or controlled cloud and data storage environments (maintained by universities, national research institutions, or partner companies) will be used. Data sharing and processing agreements will be established for data reusability, and restricted access conditions may be applied for sensitive data.

Overall, the primary objective of the initial version of the DMP is to establish a comprehensive strategy and framework for the reuse, collection, sharing, and storage of data within the REALM project. This DMP differentiates the handling and administration of data according to its identifiability, namely individually identifiable real-world data, anonymized research data, and synthetic data derived from real-world data but individually unidentifiable. We will keep updating the DMP in the project lifetime to ensure proper data management practices, enhance data availability and reuse, facilitate collaboration within REALM and with other EU projects, and meet the requirements of funding agencies (HE), research institutions, and regulatory bodies.

8 Next steps (if applicable)

This DMP will be kept updated during the project time. Based on the project needs and development, the updated information related to data, software, and other research output in REALM can be indicated and discussed in Steering Committee meeting (every three months). The next official version of DMP (Deliverable D4.3) will be delivered by 30 June 2025.