pxCall: 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.144

D4.3 Data Management Plan V2.0

Contractual Submission Date: 30/June/2025

Actual Submission Date: 30/June/2025

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	**D4.3**
Deliverable title	**Data Management Plan V2**
Type[1]	R
Dissemination level[2]	PU - Public
Work package number	WP-4
Work package leader	Chang Sun (P1-Maastricht University)
Author(s)	Chang Sun (Maastricht University) Frederic Jung (VITO), Stef Rommes (VITO) Aris Bonanos (EXUS), Pedro Lopez (EXUS)
Reviewers	Dominika Harasimiuk (University of Warsaw) Bart Elen (VTIO)
Keywords	DMPV2, Updated Data management Plan, Research data, FAIR, EHR, Standardization, Data Management

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
V0.1	02/April/2025	Draft DMP V2 creation
V0.2	11/April/2025	Initial input about data catalogue, OMOP, data access agreement
V0.3	29/May/2025	Draft version with input and updated from all partners
V0.4	19/June/2025	Complete version for internal review
V1.0	19/June/2025	Finalized version for official submission

[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

CDM – Common Data Model
COPD - Chronic Obstructive Pulmonary Disease
CT Scan - Computerized Tomography scan
DARWIN EU – Data Analysis and Real World Interrogation Network
DDL – Data Definition Language
DICOM - Digital Imaging and Communications in Medicine
DMP - Data Management Plan
EHDEN – European Health Data Evidence Network
EHR - Electronic Health Record
ETL – Extract Transform and Load
FAIR - Finable, Accessible, Interoperable, Reusable
FHIR - Fast Healthcare Interoperability Resources
GDPR - EU General Data Protection Regulation
HCTS - Human-Computer Trust Scale
HITL – Human In The Loop
KFP – Kubeflow Pipelines
OHDSI – Observational Health Data Sciences and Informatics
OMOP - Observational Medical Outcomes Partnership
REALM - Real-world-data Enabled Assessment for heaLth regulatory decision-Making
RWD – Real-World Data
SOAP note - Subjective, Objective, Assessment and Plan (SOAP) note
SUS - System Usability Scale
TAM - Technology Acceptance Model
BMI – Body Mass Index
ICU – Intensive Care Unit
DoB – Date of Birth

Executive Summary

This updated Data Management Plan (DMP V2.0) presents the updated status and evolution of data management practices in the REALM project, reflecting developments and lessons learned since the initial DMP (D4.2). The core objectives remain the same: to ensure data integrity, enable secure and ethical data sharing, and promote compliance with European regulations and Horizon Europe requirements. This version integrates updated inputs from all demonstrators, along with progress and refinements from WPs 2, 3, 5 and 6.

REALM aims to build a trustworthy, inclusive platform for evaluation and certification of AI-enabled healthcare software by combining real-world, synthetic, and anonymized data with a standardized technology stack. Over the past 30 months, significant progress has been made in establishing the REALM architecture, integrating tools, and building the data repositories that support testing and validation workflows. Data management has played a central role in ensuring that these activities remain aligned with FAIR principles and regulatory expectations.

This DMP V2.0 documents the types and volumes of data collected or generated, including updates to data provenance, metadata standards, and data quality practices. We have classified and managed three categories of data:

1. personal real-world data from demonstrators under strict access controls.
2. anonymized datasets from biobanks and trusted repositories; and
3. synthetic data generated to support algorithm development and benchmarking.

This DMP V2.0 captures the REALM project's commitment to responsible and transparent data management. It ensures that data assets remain FAIR, secure, and privacy-preserving, supporting the immediate project goals and broader scientific and regulatory communities in healthcare AI.

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

Demographic data, electronic health records (EHR), medical imaging data, medical sensor data, genomic data, and questionnaire data has been collected and re-used. REALM started with requesting the access to publicly available data from multiple data repositories including but not limited to MIMIC Clinical Database (https://physionet.org/content/mimiciv/3.1/), UK Biobank (www.ukbiobank.ac.uk), and Kaggle Data Platform. Next, bilateral agreements have been set up, and access to use patients' pseudonymized data from hospitals, clinics, research institutes, and other healthcare providers in the Netherlands, Belgium, and Greece has been requested (refer to Figure 4: REALM Component Architecture in the REALM Grant Agreement). Since the first version of the DMP (D4.2), five demonstrators provided their data description and test data for evaluating the REALM infrastructure. The five demonstrators are:

1. **Dune AI,** from Maastricht University, performs automatic detection and volumetric (3D) segmentation of non- small-cell lung cancer (NSCLC) tumours in CT scans. Dune AI re-uses existing patients' CT Scan data and their demographic information and will collect System Usability Scale (SUS) evaluations data, and qualitative feedback from clinicians under WP6.
2. **PGX2P**, from VITO, tests a set of germline genetic variants representing a new model for employing stratified medicine in practice to predict disease and therapeutic response to drugs used. PGX2P has genotyping, EHR, and demographic data from 100 individuals. However, due to the high sensitivity of this data, PGX2P cannot share the original dataset. Instead, it re-uses a comparable pharmacogenomics dataset from the Centres for Disease Control and Prevention's (CDC) Genetic Testing Reference Materials Coordination Program (GeT-RM) for development and evaluation purposes within the REALM infrastructure.
3. **STAR**, from Liege University, is a model-based decision-support system to predict blood glucose ranges for any insulin and nutrition intervention, and any forward interval out to 6-hours and provide recommendations of insulin doze, maximize time in band and limit the risk of glucose below normal to less than 3%. STAR re-uses patients' data including their demographic and EHR data including diabetes status, insulin rates, nutrition rates, glucose, and other data for patient response to insulin.
4. **ASCOPD**, from TRAQBEAT, is a tool to predict the needed medical attention for COPD patients. It re-uses the existing patients' data including demographic and EHR data from admission and discharge events and generate new heart-related data using TRAQBEAT sensors.

[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

5. **COPowereD**, from COMUNICARE, is a generic app to follow up with COPD patients and predict the detection of COPD exacerbations. COPowereD re-uses and will collect new observation and vital sign data from patients with chronic obstructive pulmonary diseases from hospitals.

REALM re-used sensitive and individual datasets which cannot be accessed and shared by third parties or can only be accessed and shared under restrictions. In case of restricted data access, we identified and re-used data from publicly accessible database such as MIMIC Clinical Database from Physionet, GeT-RM, Kaggle platform for each demonstrator. These data were used for testing and experimenting the development of REALM infrastructure, ensuring the REALM architecture can be built without a delay in the project timeframe. Please find the list of equivalent public datasets for the demonstrators below.

Table 1. Re-used existing public dataset for each demonstrator

Demo	Disease	Equivalent public dataset	License	Dataset Link
Dune AI	Lung Cancer	Lung-PET-CT-Dx | A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis	CC BY 4.0	www.cancerimagingarchive.net/collection/lung-pet-ct-dx/
PGX2P	-	Genetic Testing Reference Materials Coordination Program (GeT-RM)	Not specified	www.cdc.gov/lab-quality/php/get-rm/reference-materials.html
STAR	Diabetes	MIMIC Clinical Database	PhysioNet Credentialed Health Data License 1.5.0	physionet.org/content/mimiciii/1.4/
ASCOPD	COPD	Kaggle – COPD Patients Dataset	CC0 1.0	www.kaggle.com/datasets/prakharrathi25/copd-student-dataset
COPowereD	COPD	A machine learning approach to triaging patients with chronic obstructive pulmonary disease	Not specified	https://pmc.ncbi.nlm.nih.gov/articles/PMC5699810/#notes1

Since the initial version of the Data Management Plan V1.0 (D4.2 submitted in M6), we have actively worked on implementing data access agreements (D4.5, M9) that can be signed by the REALM national nodes. The agreement template has been iteratively updated to reflect evolving requirements, notably to 1) incorporate the access modalities defined in D4.1 (M31), and 2) include a clear and comprehensive research plan annex. Special emphasis has been placed on engagement with the Greek REALM node and the TraqBeat (TB) use case, which is also being developed in Greece. Discussions are ongoing with several Greek data pools (detailed in Table 2), which are federated under the OHDSI Greek node, with Pantelis Natsiavas serving as the contact point. These data partners, being part of the OHDSI community, are already compliant with the OMOP Common Data Model, enabling immediate technical readiness for use once data access agreements are signed. The data partners already shown interest in joining the project, the outcomes of these negotiations, including finalized agreements with the REALM Greek node, will be documented in D4.1.

Table 2: OHDSI Greek node registered data partners

Organisation	Type	Email
General Hospital of Kavala - EHR	General hospital (public)	pliroforiki@kavalahospital.gr

Papageorgiou General Hospital of Thessaloniki - EHR	General hospital (public)	pmo@papageorgiou-hospital.gr
Hygeia Hospital - EHR	General hospital (private)	info@hygeia.gr
Greek CLL dataset (hosted by INAB|CERTH)	Dataset collected by various haematological centres across Greece	eva.minga@certh.gr
Digital Health Solutions	Diagnostic lab data	timo.bilalis@bioiatriki.gr

1.2 Types and formats of data

Table 3. An overview of data types, formats, and elements in REALM

Data Types	Formats	More information about data
Demographic	Structured data	Age, sex, ethnicity, population, superpopulation, 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 (SUS) data	Likert Scale / comments (text data)	Likert scales and potentially written comments from the system usability scale assessments.
Qualitative Feedback / Technology Acceptance Model (TAM) data	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews in the context of the technology acceptance model.
Human-Computer Trust Scale (HCTS) data	Likert Scale / comments (text data)	Likert scales and potentially written comments from the human-computer trust scale assessments.
Usability of potential new features evaluations	Likert Scale / comments (text data)	Likert scales and potentially written comments from the assessment of the usability of potential new features for the demonstrator. (Limited to PGX2P demonstrator)

Table 4. Descriptions of datasets from five demonstrators in REALM

Demo	Data Types	Formats	Elements	Inclusion
Dune AI	Demographics & EHR	Structured data	Age, sex, disease, and other clinical information	Less 5mm thickness - images, optional

	CT Scans	Image data	Medical images of patients' lungs used by the DuneAI for tumor segmentation in DICOM (Digital Imaging and Communications in Medicine) format.	is contrast agent, hardware Kernel.
	SUS evaluations	Likert Scale / comments (text)	Likert scale data and potentially written comments from the system usability scale assessments.	
	Qualitative Feedback / TAM evaluations	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews in the context of the technology acceptance model.	
	HCTS evaluations	Likert Scale / comments (text data)	Numerical/categorical scores and potentially written comments from the human-computer trust scale assessments.	
PGX2P	Demographics & EHR	Structured data	Sex, age, ethnicity, diagnosis, and medication data	Must have genotyping, diagnosis and medication data
	Genomics	Sequencing data	Genotyping, Exome Sequencing, Whole Genome Sequencing	
	Genotyping	Sequencing data	Collect from a small cohort for validating the application.	
	SUS evaluations	Likert Scale / comments (text data)	Likert scale data and potentially written comments from the system usability scale assessments.	
	Qualitative Feedback / TAM evaluations	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews in the context of the technology acceptance model.	
	HCTS evaluations	Likert Scale / comments (text data)	Numerical/categorical scores and potentially written comments from the human-computer trust scale assessments.	
	Usability of potential new features evaluations	Likert Scale / comments (text data)	Likert scale data and potentially written comments from the evaluation of potential new features.	
STAR	Demographics & EHR	Structured data	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	Measurements in ICU in emergent treatment, measurement duration is 1-3 hours.

	SUS evaluations	Likert Scale / comments (text data)	Likert scale data and potentially written comments from the System Usability Scale assessments.	
	Qualitative Feedback / TAM evaluations	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews in the context of the technology acceptance model.	
	HCTS evaluations	Likert Scale / comments (text data)	Numerical/categorical scores and potentially written comments from the human-computer trust scale assessments.	
ASCOPD	Demographics & EHR	Structured data	DoB, address, smoking, allergies, vaccination status, BMI, medical history, Lab tests on admission and discharge, physical activity, smoking status, weight.	Patient with diagnose COPD admitted at ICU
	Heart monitoring	Time-series data	Vital signs: ECG, heart rate, heart rate variability, blood pressure, respiratory rate, SpO2, skin temperature	
	Chest X-Ray	Image data	X-Ray images	
	SUS evaluations	Likert Scale / comments (text data)	Likert scale data and potentially written comments from the system usability scale assessments.	
	Qualitative Feedback / TAM evaluations	Text data	Written or transcribed spoken feedback from clinicians, gathered from think-aloud sessions and semi-structured interviews in the context of the technology acceptance model.	
	HCTS evaluations	Likert Scale / comments (text data)	Numerical/categorical scores and potentially written comments from the human-computer trust scale assessments.	
COPowereD	Demographics & EHR	Structured data	Age, sex, disease, symptoms, and other clinical information	Patient with diagnose COPD
	Lung monitoring	Time-series data	Vital signs	
	SUS evaluations	Likert Scale / comments (text data)	Likert scale data 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	

	TAM evaluations		sessions and semi-structured interviews in the context of the technology acceptance model.	
	HCTS evaluations	Likert Scale / comments (text data)	Numerical/categorical scores and potentially written comments from the human-computer trust scale assessments.	

1.3 Purpose of data generation and re-use

REALM re-used the existing data from health data repositories and sources with public or on-request access to develop and test the REALM platform. Under WP4, we created the data catalogues to aggregate and harmonize existing data from different RWD sources across multiple institutions and public repositories. We designed an evaluation workflow that extracts datasets out of those data sources to evaluate AI components in REALM. The required information for the extraction of the evaluation datasets is fully documented, reproducible and based on cohort definitions, demographics, inclusion criteria (metadata) and model input (features) mapped to the OMOP CDM (The Observational Medical Outcomes Partnership Common Data Model).

During evaluation, the extracted datasets are stored temporarily within the evaluation workflow solely for execution purposes. Once the query builder is invoked, the dataset is passed into the entry components of the evaluation pipeline as input. These data elements, known as Artifacts, are consumed and/or produced by components within the Kubeflow Pipelines (KFP) framework. To maintain stateless execution, caching is disabled for all components in the evaluation pipeline. Consequently, neither the dataset artifacts nor any intermediate artifacts, such as metrics, models, or internal variables, are retained after execution.

In addition to re-use existing data, REALM also deployed generative models to generate synthetic data which is statistically and structurally similar to the real data. The purpose of synthetic data generation is to facilitate REALM to be still able to evaluate AI models that are trained on restricted data and evaluate AI models on a synthetic population that is different from the trained data. The generative models are installed in the REALM infrastructure. Only if the user indicates to use the synthetic data to evaluate the AI models, the generative models will be trained to generate synthetic data. The generated synthetic data and trained generative models are not stored in the evaluation workflow but in a node of the federated data cloud.

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

The size of the generated or re-used data in the overall project is about 500GB 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.

1.5 Origin/provenance of the data

The origins/provenances of data are the data sources - data owners' repository, and hospitals and research institutes that collect the original data. In the evaluation phase, the origins/provenances 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

The data is 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) identified the stakeholders and actors for REALM and analysed their user requirements, security and privacy requirements. The collected and generated patient data from REALM will be useful for health researchers in Lung Cancer, diabetes, COPD, and genomic research areas. The collected questionnaire data will be useful for regulatory authorities and policy makers and the data and model evaluation results data will be useful for certification authorities and medical device manufacturers.

2 FAIR Data

2.1 Making data findable, including provisions for metadata

2.1.1 Persistent identifiers and metadata of data

The public datasets used in this project obtained from such databases as MIMIC and Cancer Imaging Archive keep their original persistent identifiers. For the metadata of the datasets, all data sources connected to the REALM nodes are systematically catalogued using standardized metadata, ensuring they are discoverable, reusable, and well-described in accordance with FAIR principles. Each dataset integrated into the REALM infrastructure is documented using structured information drawn from the Common Data Model “CDM_SOURCE table”, which plays a central role in describing the origin and technical details of each OMOP CDM instance. The metadata includes:

- **Data source of origin:** Captured in the *cdm_source_name* and *source_description* fields, these values provide the name of the dataset and a free-text description of its provenance, collection context, and any defining characteristics.
- **ETL reference documentation:** The *cdm_etl_reference* field points to the documentation, protocol, or publication describing the Extract-Transform-Load (ETL) process applied to convert the source data into OMOP CDM format.
- **CDM version:** Recorded in the *cdm_version* field, this value specifies which version of the OMOP CDM was used to structure the dataset (V5.4). This is essential for compatibility with analytical tools, vocabulary mappings and integration with third-party applications.
- **Vocabulary version**: This refers to the specific release of the OMOP standardized vocabulary used in the database (v20240830). This version can be traced back to the ATHENA platform or requested directly from the REALM node (REALM-hosted ATHENA under development).
- **Clinical domains covered**: Not stored directly in the CDM_SOURCE table, information on the data domains populated (e.g., conditions, drugs, measurements) can be derived from the dataset and is reported in the model card.

- **Distribution of training sets (for synthetic generation)**: Summaries and stratifications of training data used for synthetic data generation are documented alongside the metadata and included in the model card for transparency.

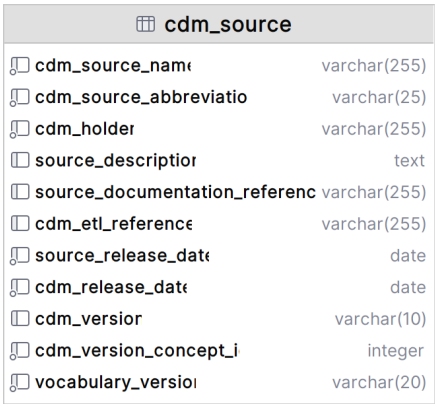

Figure 1: Structure of the CDM_SOURCE table in OMOP CDM *– Exported UML table of the CDM_SOURCE used in REALM within all OMOP data sources. This table captures the metadata about the data source itself, including its name, description, CDM version and the vocabulary version in use. This table can be use and query for getting the provenance and context of the clinical data stored in the OMOP instance and thus, understand further the clinical context of the evaluation dataset.*

Finding cohorts is a fundamental part of the REALM infrastructure, enabling the selection of patient groups that meet specific criteria for AI model evaluation. A cohort is defined using logical expressions such as “male adult patients diagnosed with COPD.” To streamline and standardize this process, REALM provides a user-friendly web interface within the RIANA dashboard, allowing users to define cohorts in a clear and structured way. Once the logic of a cohort is finalized, it is assigned a unique identifier: the *cohort_definition_id*. This ID acts as a persistent internal reference, making the cohort definition reusable, traceable, and executable across different datasets and institutions. The detailed definition is stored in the *COHORT_DEFINITION* table, which includes not only the ID and name of the cohort but also the entire logical expression in a machine-readable and comprehensive JSON format. This format ensures the definition is fully documented and exportable, which is essential for reproducibility, especially when repeating evaluations on other datasets or in new locations. This JSON description is also embedded in the model card and later published publicly via blockchain, enhancing transparency and trust.

Cohort data are temporarily stored in the *COHORT* table, which contains the actual list of patients matching the criteria. Each patient is identified by a *subject_id*, corresponding to the *person_id* used in the clinical data tables. This allows real clinical data to be linked to the cohort via the cohort_definition_id, forming the basis for further data extraction and feature generation needed for creating evaluation datasets.

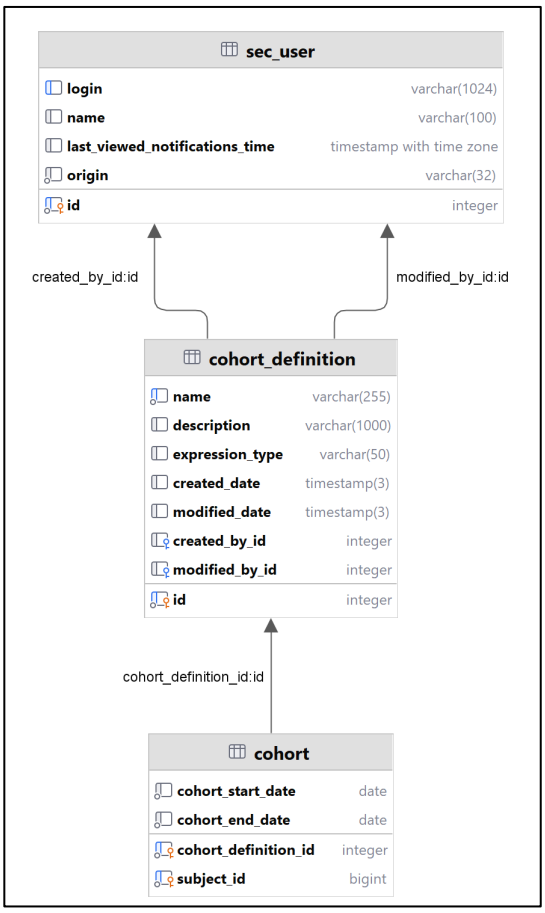

Figure 2: A cohort system diagram illustrating the structure and relationships between three essential tables for the cohort creation and generation. This diagram includes field names and types and an overview of foreign primary keys defined by the OHDSI v5.4 DDLs.

2.1.2 Keywords in metadata

The REALM standardized OMOP data catalogue is annotated using controlled vocabularies, specifically the OMOP standardized vocabulary based on unique concept IDs. These concept IDs serve as metadata keywords, enabling semantic clarity, machine-readability, and interoperability across data sources within the REALM infrastructure. The concept IDs embedded in the metadata describe the key characteristics of each dataset and are aligned with the requirements of the demonstrators, specifically the features expected as input by the AI models under evaluation.

The metadata keywords reflect several dimensions, including:

- OMOP domains: such as *condition_occurrence*, *observation*, and *measurement*, which define the type of clinical information available.
- Data modalities: including flattened EHR data prepared as input features for models, as well as support for genomic (G-CDM) and imaging (R-CDM) extensions.
- Linked ontologies and terminologies: such as SNOMED CT, LOINC, ATC, and others integrated into the OMOP vocabulary.

This use of standardized concept identifiers not only ensures internal consistency across the REALM catalogue but also enables future integration with external FAIR-compliant infrastructures, such as DARWIN EU, EHDEN, or other federated data networks.

2.2 Making data accessible

2.2.1 Trusted data repository

The project deposits the generated synthetic data that is evaluated as “non-personal data” in an open-access repository such as Figshare or Physionet. 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” is managed and maintained at their source site. 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. 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 will be published and stored in the trusted repository when the project is finished.

2.2.3 Data identifiers in the repository

The project only uses the data repositories and services such as OMOP, Physionet that can provide the persistent identifiers and assign a unique persistent identifier to a digital object.

2.2.4 Data availability

REALM project includes existing health research data from biobanks and data repositories, generated synthetic data based on real-world patient data, and actual real-world patient data. REALM keeps the existing published research data at their original sites and indicate the sources, request time, and procedures. The anonymized real-world data and synthetic data that are not individually identifiable can be provided accessible for research.

Data sharing and data processing agreements, confidential agreements are set up by the data controllers based on EU member states’ data protection laws and GDPR to support the restricted data access. The purpose of data access, the method of data process, the requirement of informed consent, and other essential conditions are established in these agreements. The original real-world data that are individually identifiable cannot be shared with any third parties without consent from individual data subjects.

2.2.5 Potential embargo

An embargo may be applied to give time to publish or seek protection of intellectual property such as 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 generated synthetic data and anonymized real-world data which are individually unidentifiable can be accessible through a free and standardized access protocol. Only the metadata of the sensitive and individually identifiable data will be accessible through a free and standardized access protocol. The sensitive and pseudonymized real-world data which are individually unidentifiable could potentially be accessible through a secure, standardized, and controlled environment. However, the data source decides on the data access control.

2.2.7 Identity of data requester

The project does not host or control any sensitive and restricted data use/sharing. If any requester intends to access to the data from the demonstrators or the data repositories, they must go through the data application from each data source. The identity of the person accessing the data will be ascertained and validated by the data sources.

2.2.8 Data access committee

The project does not host or control any sensitive and restricted data use/sharing. Therefore, the project does not form a data access committee to process the data request from other researchers.

2.2.9 Metadata availability

The metadata of all the datasets will be made 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.

2.2.10 Availability period

The data will remain available and findable for 10 years and metadata will remain available even after data is no longer available.

2.2.11 Software requirements of accessing data

Accessing and interacting with data directly from REALM is via the RIANA dashboard. The users can indicate the cohort definition, execution, and standard analytics on the web-based interfaces or third-party applications using API calls, without the need for direct database access. The version of the WebAPI documentation currently used in the REALM test node is accessible to the REALM consortium (tested on our EC2 test instance under port 8081) and provides a detailed description of all endpoints connecting to OMOP data sources and vocabulary schemas. This API endpoint documentation can be setup by notional nodes and open on demand to developers, analysts, data engineer or data partners.

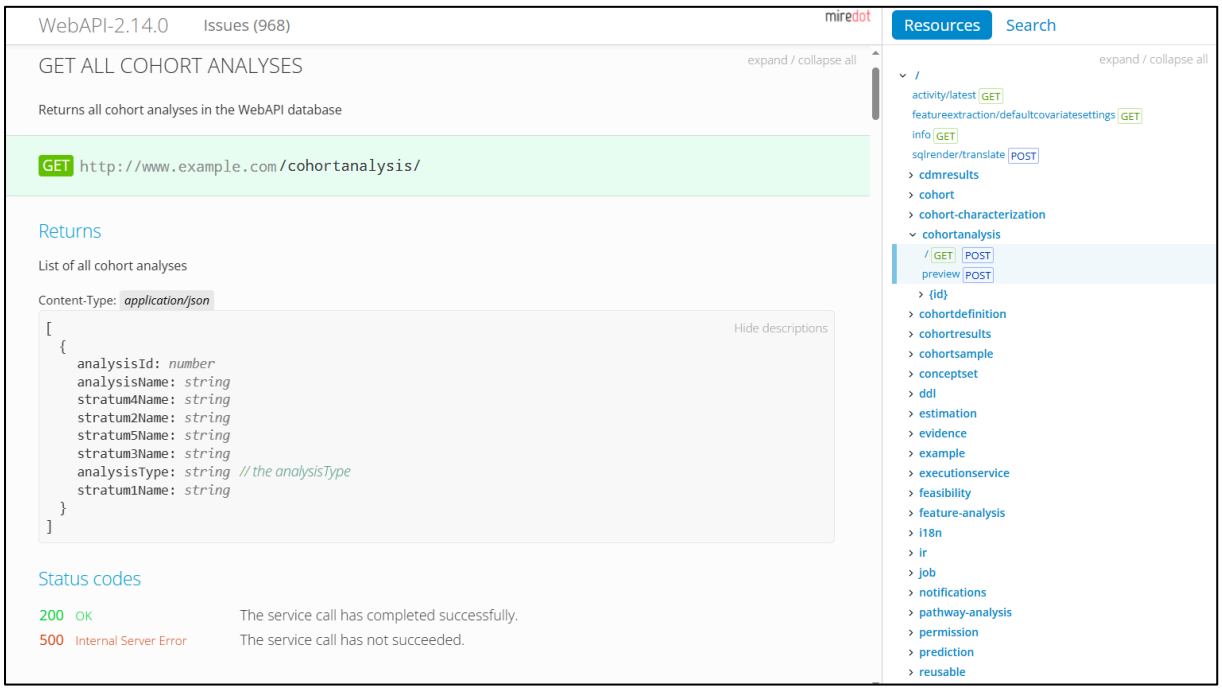

Figure 3: WebAPI Endpoint documentation for REALM infrastructure – This figure presents the online documentation that we generated of the current OHDSI WebAPI version used in REALM infrastructure. It provides a comprehensive overview of all available endpoints, including those for cohort generation, definition, and analysis. The documentation was automatically generated from the API using Miredot and packaged into a container for seamless deployment.

Registered users and organizations can define and execute cohorts, get analytics (Heracles) and perform initial data sources check via the RIANA dashboard. These operations can be initiated in ATLAS or using API endpoint.

The minimal backend infrastructure requires:

- A PostgreSQL database hosting the OMOP CDM instances
- Access to OHDSI webAPI
- A user interface such as Atlas or the RIANA dashboard

In the evaluation workflow, REALM incorporates a dedicated query builder component designed to extract data from the OMOP CDM for use in downstream components, including those supporting AI-based synthetic data generation, training, and evaluation. Upon receiving a request (claim) from the RIANA dashboard via the middleware API, the query builder interprets the user-defined cohort specifications to query the OMOP-formatted database. The resulting data is transformed into a standardized RAW format, which serves as input for various stages of the AI pipeline, including data pre-processing, evaluation, and explainable AI modules.

2.3 Making data interoperable

Interoperability is a foundational principle of the REALM infrastructure, ensuring that data and tools can be reused, shared, and understood across institutions, systems, and regulatory environments. REALM is built using the internationally recognized OMOP Common Data Model (CDM), which standardizes both the structure and semantics of healthcare data.

A clear example of this interoperability is found in the cohort definition process. Cohorts created within the REALM platform (especially through the RIANA dashboard) follow the same logic and format used by the broader OHDSI ecosystem, including tools like Atlas. These cohort definitions can be exported and reused across OHDSI environments, enabling researchers and regulators to replicate evaluation logic on different OMOP datasets and conduct consistent observational studies, as OMOP was originally intended for. This capability supports federated analyses, where multiple sites run the same queries locally while preserving data privacy, and then share aggregated results.

Beyond structural compatibility, REALM also ensures semantic interoperability. All clinical concepts used in datasets are mapped to standard concept IDs, making them interpretable across institutions and software platforms. These concepts are linked to widely adopted terminologies such as SNOMED CT, LOINC, and RxNorm, and can be aligned with exchange standards like HL7 FHIR, ensuring compatibility with a broad range of data formats and institutional systems.

2.3.1 Data and metadata vocabularies

REALM ensures interoperability by standardizing both the structure of the databases and also the semantics of the data by using the OMOP Common Data Model (CDM) version 5.4. Each entry in our data catalogue corresponds to a distinct OMOP data source, typically hosted in a PostgreSQL database instance. These data sources originate from different hospitals, networks, and clinical settings, and are harmonized through a consistent schema and vocabulary framework.

Table 5: List of Vocabularies Used in REALM (OMOP CDM v5 – Vocabulary Version v20240830) – This table displays the list of standard and extension vocabularies included in the REALM infrastructure. Each entry includes the vocabulary ID, the associated CDM version and the code used in the OMOP vocabulary system and Athena.

ID	CDM	Code (cdm v5)	Name
146	CDM 5	OMOP Genomic	OMOP Genomic vocabulary of known variants involved in disease
145	CDM 5	OncoKB	Oncology Knowledge Base (MSK)
144	CDM 5	UK Biobank	UK Biobank (UK Biobank)
141	CDM 5	Cancer Modifier	Diagnostic Modifiers of Cancer (OMOP)
128	CDM 5	OMOP Extension	OMOP Extension (OHDSI)
87	CDM 5	Specimen Type	OMOP Specimen Type
82	CDM 5	RxNorm Extension	OMOP RxNorm Extension
44	CDM 5	Ethnicity	OMOP Ethnicity
34	CDM 5	ICD10	International Classification of Diseases, Tenth Revision (WHO)
13	CDM 5	Race	Race and Ethnicity Code Set (USBC)
12	CDM 5	Gender	OMOP Gender
8	CDM 5	RxNorm	RxNorm (NLM)
6	CDM 5	LOINC	Logical Observation Identifiers Names and Codes (Regenstrief Institute)
1	CDM 5	SNOMED	Systematic Nomenclature of Medicine - Clinical Terms (IHTSDO)

The structure of each OMOP database strictly follows the OHDSI Data Definition Language (DDL) specifications maintained by the OHDSI community. The database schema includes a full set of clinical

data tables (e.g., *person*, *condition_occurrence, drug_exposure, measurement, observation,* etc.) that store patient-level information. These tables are used to describe health events and are directly populated during ETL from source datasets. To facilitate deployment, we provide a dockized database setup that can rapidly create an OMOP PostgreSQL instance based on the official DDLs. Alongside with the OMOP data sources, we use standardize vocabulary which allow the semantic alignment across all the datasets. In our implementation, the OMOP vocabulary tables are stored in a dedicated schema, separate from the clinical data. These vocabulary tables enforce all OMOP data sources to follow the standardized vocabulary by using foreign key constraints that link clinical fields (e.g., *condition_concept_id, drug_concept_id*) to standardized concepts. And thus, ensure that all concept codes referencd in the clinical table are valid and standardized.

The vocabulary schema consists of 9 tables that can be requested from Athena, some of them are explained in the table below. The vocabulary tables were retrieved from OHDSI Athena (version v20240830) and manually imported into our REALM vocabulary schema. The location of the vocabulary can optionally be indicated to our dockize database setup to load the vocabulary once the tables are created.

Table 6: Vocabulary schema of tables and definitions.

Table	Definition
concept	Contains all standardized concepts and all medical terminologies and concept are associated with a unique concept_id
vocabulary	Lists all the coding systems and ontologies that have been aggregated to create the OMOP vocabulary (e.g., SNOMED, LOINC, ICD). This can be used to map the source codes to the right standard concept in the OMOP vocabulary.
domain	Defines in which clinical table in the OMOP CDM a concept can be use (e.g. Condition, Drug, Measurement).
concept_relationship	Describes how two concepts are related, using relationship (e.g., "Maps to", "Is a"). Can be used for creating concept sets.
concept_ancestor	Hierarchy table that keeps track of the ancestor and descendant relationships between concepts (e.g., all descendants of a disease category, severity level, disease stages…).

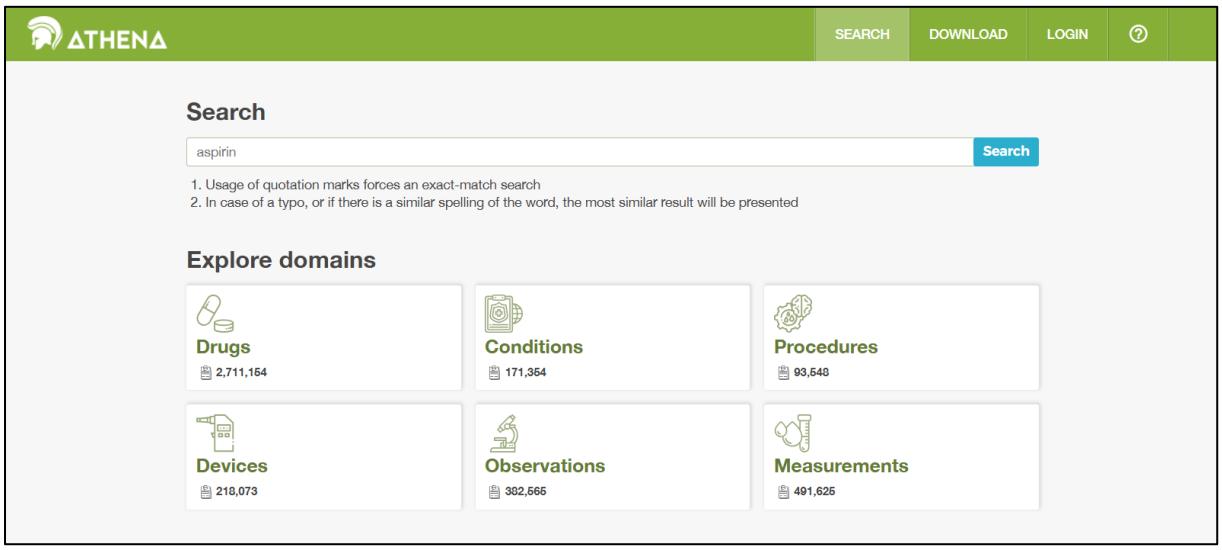

Figure 4: ATHENA Landing Page – OMOP Vocabulary Search Interface – this figure shows the landing page of the REALM local version of Athena, the vocabulary search engine developed by OHDSI. Athena allows users to search, explore and request standardized vocabulary tables.

2.3.2 Linking data with other resources

REALM enables cross-modality data integration, linking diverse medical data types such as clinical records, imaging (DICOM), and genomics. This is achieved by leveraging OMOP CDM extensions such as G-CDM (Genomics-CDM) and R-CDM (Radiology-CDM), which are specifically designed to support non-traditional clinical data. In this architecture, essential metadata (e.g., age, gender, ethnicity, and clinical observations) is harmonized and stored in OMOP's core clinical tables. This standardized representation allows data from different domains to be linked and queried in a coherent, consistent way. The goal is to connect all medical records and associated metadata through the OMOP CDM, creating a holistic view of each patient's data across sources and modalities. This integrated approach not only supports comprehensive AI model evaluations within the REALM platform but also enables reuse of the data with external tools and platforms.

2.4 Increase of data re-use

2.4.1 Data documentation for re-use

To support transparent and reproducible AI model evaluation, REALM generates a model card for each AI model assessed. These model cards provide a detailed description of the target patient population, the data used for evaluation and the conditions under which the model was tested. More importantly, each model card also documents how the evaluation dataset was constructed, and explain in detail the cohort definition and the domains used. As all the features are mapped to the OMOP data model, the structure and the semantic of the data are explicit and standardized avoiding misinterpretation and giving a detail clinical context based on metadata.

To ensure trust and traceability, each model is published on the blockchain, making the documentation immutable and publicly accessible. This allows any authorized stakeholder with access to the model

care to fully understand how the data was used, the origin of the data, under what conditions it can be used and what subset of the OMOP data catalogue was used. Anyone with access to the source data can re-generate on demand the evaluation datasets or apply the same cohort definition to any OMOP data sources that have described features.

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, it's not appropriate to make these data freely available in the public domain.

2.4.3 Data re-usability during and after the project

In REALM evaluation workflow, the evaluation datasets are only stored temporally for execution purpose. They are removed after the model evaluation phase, but can be re-generated on demand, using the same documented extraction logic (model card) and source data (data catalogue).

For the reused public datasets in the project, we provided data description and data access link to the data sources for other researchers to reuse. For the generated synthetic patient data which is seen as individually unidentifiable data, it can be used by third parties under data request of the original data. The trained generative models apply the same requirements as the synthetic data itself. Both the generated synthetic data and the trained generative models are stored in one of the REALM federated data nodes (BE, NL, GR). The sensitive and individually identifiable patient data will not be re-used by the third parties without data request to the original data request.

2.4.4 Data provenance

For re-used data that are from repositories, we keep their original data provenances in our project and keep them updated. For real-world data that are collected during the project, the project documents the provenance of the data such as where data originate, where they were collected, by whom, and for what reason. To keep tracking the data versions in REALM project, we log the updates and changes of the data in the model cards. The data provenance can be stored and archived for citable reproducibility. For the synthetic data that is generated from the real-world data, we document the metadata of the training data, the structure and hyperparameters of the generative models.

2.4.5 Data quality assurance

The project established a 26-dimension data quality framework to evaluate the data quality based on different types of data [Deliverable 4.4. Standard Data Quality Framework]. The framework covers data accuracy, completeness, consistency, reliability, and adherence to data collection protocols. We conducted the qualitative and quantitative evaluation based on these 26 dimensions on the equivalent datasets of the five demonstrators to assess the data quality. We document the reviewed data, results, identified data quality issues in the Deliverable 4.4.

3 Other research outputs

3.1 FAIR Principle for other research outputs

The research outputs in REALM project are 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 and 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). There are two sensitivity levels of the outputs that are indicated in the REALM grant agreement Page 22-28 – public (PU) and sensitive (SEN). For the outputs that can be public, the code of the developed infrastructure and models are open source, uploaded to public Github repositories, and maintained during and after the project. A list of published repositories from finished tasks:

- Standard Data Quality Framework (T4.2): https://github.com/Metamind-Innovations/Standard_Data_Quality_Framework
- XAI tools for REALM use cases (T3.3):
  - https://github.com/Metamind-Innovations/realm_task_3_3_implementation_xai_uc1
  - https://github.com/Metamind-Innovations/realm_task_3_3_xai_uc2
- Post-Market for REALM Use Cases 1 (T5.1):
  - https://github.com/Metamind-Innovations/realm_task_5_1__postmarket_uc1
  - https://github.com/Metamind-Innovations/realm_task_5_1_postmarket_uc2

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

4 Allocation of resources

4.1 Estimated costs for making data and research output FAIR

Our strategic design choices and OMOP workflow in the project enhance making our data and research output FAIR. REALM adopts a modular OMOP-based workflow and a reusable ETL scaffolding, as detailed in D4.1 (M31). This approach significantly reduces the effort required to transform new datasets. Once a structural mapping is available, only the SQL transformation scripts need to be implemented, minimizing the workload for each new data source. As REALM targets data pools that are already part of an OMOP federated network, meaning their data is already standardized and aligned with the FAIR principles. This reduces the harmonization costs and investment. Research outputs are documented using standardized model cards, detailing model requirements, target populations, evaluation setups, and results. This ensures replicability and facilitates the assessment of new data sources against model input criteria.

4.2 Resources

We allocated 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 storage and data quality. T4.1 and T4.3 focused on making the data and research outputs FAIR in the project. The costs related to data requests, storage, transferring, and security during the project time are 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 people 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. 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 and the generated datasets. For the synthetic data generated on the real data from demonstrators, the demonstrators have the responsibility for the synthetic data. For the synthetic data generated on the public datasets, the synthetic data is uploaded to data repository and accessible for people who have access to the original data.

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 re-used research data from data repositories are maintained by their origins. The generated synthetic data be preserved for 10 years for scientific research use. The original 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

Sensitive and individually identifiable real-world data stay at their sources in a secure encrypted storage. The security standards of the data sources will be applied in this case. The non-sensitive and unidentifiable synthetic data is stored in a trusted research data repositories (Figshare, Physionet). The security standards of the repository will apply in this case.

5.2 Data storage and preservation and curation

Due to the different nature of data used in the project, we stored non-sensitive synthetic data in open-access and public research repository. Non-sensitive and anonymized data that are individually unidentifiable will 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 and storage at the data source site.

6 Ethics

6.1 Possible ethical or legal issues on data sharing

The real-world patient data in the project 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. The potential re-use of the data in the future may be limited by the requirements resulting from the GDPR. All the individual personal data are and will be collected with consent and will be anonymized before sharing to mitigate the risk of privacy breach. 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. Therefore, such data is kept with the restricted access.

6.2 Informed consent

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

7 Conclusion

This updated Data Management Plan (DMP Version 2) reflects the current status of data handling, access, and reuse in the REALM project and builds upon the initial version submitted at the start of the project. It is aligned with the requirements of the Horizon Europe Model Grant Agreement and the associated guidance in the Grant Agreement.

REALM manages both re-used and newly collected sensitive real-world data under strict ethical and legal frameworks, with all necessary data access agreements and GDPR-compliant procedures in place. The project leverages the OMOP Common Data Model, medical terminologies and vocabularies to ensure interoperability, standardization, and FAIR data principles across all nodes of its federated architecture.

New elements in this DMP version include updated data sources, enhanced metadata handling, integration of synthetic data generation workflows, and improved documentation practices via model cards. These improvements ensure reproducibility, transparency, and traceability of AI model evaluation processes.

REALM also introduces refined mechanisms for metadata publication, long-term preservation, and public accessibility of non-sensitive datasets, while maintaining strict safeguards for personal and identifiable health data. Resources, responsibilities, and cost allocations have been clearly defined to support FAIR data throughout the project lifecycle.

This DMP Version 2 will be maintained and revised as needed to capture further developments until the end of the project, ensure regulatory compliance, and support the long-term impact and reusability of REALM's data and research outputs.