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. D4.5

Authorized access agreements with European Medical Data Pools and Digital Twins

Contractual Submission Date: 30/09/2023

Actual Submission Date: 30/09/2023

Responsible partner: VITO

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

Deliverable number	**D4.5**
Deliverable title	**Authorized access agreements with European Medical Data Pools and DT**
Type[1]	R
Dissemination level[2]	PU
Work package number	WP4
Work package leader	Vito - JUNG Frederic
Authors	**JUNG Frederic (VITO)** **ERTAYLAN Gökhan (VITO)** ELEN Bart (VITO) WOUTERS Birgit (Maastricht University) SINIOSOGLOU Ilias (METAMIND) UROVI Visara (Maastricht University) HARASIMIUK Dominika (University of Warsaw) IBRAHIM Mahmoud (Maastricht University) SUN Chang (Maastricht University)
Keywords	Data pools, Demonstrators, Digital Twin, Access agreements

Funded by the European Union. Views and opinions expressed are 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	30/09/2023	First version, ready for 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

BSC – Blockchain Smart Contracts

DICOM – Digital Imaging and Communications in Medicine

DMP – Data Management Plan

DT – Digital Twin

EDITH – Ecosystem For Digital Twins in Healthcare in Europe

EHDS – European Health Data Space

EHRs – Electronics Health Records

EU – European Union

GDPR – EU General Data Protection Regulation

OHDSI – Observational Health Data Sciences and Informatics

OMOP – Observational Medical Outcomes Partnership

OpenAIRE – Open Access Infrastructure for Research in Europe

PGX2P – Pharmacogenomics passport (VITO Demonstrator)

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

RWD – Real-World Data

RWE – Real-World Evidence

SDG – Synthetic Data Generator

UM – University of Maastricht

VITO – Vlaamse Instelling voor Technologisch Onderzoek (Flemish Institute for Technological Research)

VPHi – Virtual Physiological Human Institute

Executive Summary

Europe's diverse medical data sources offer significant potential for research, innovation, and decision-making, yet accessing them raises ethical, legal and privacy concerns that need to be establish through robust data access agreements. This deliverable builds upon the effort made on the Data Management Plan (DMP), outlines the data providers, initial strategies, processes, and resources employed to identify relevant data pools and implement access agreements with European medical data pools and digital twin community.

This first version of the data access agreement template provides a comprehensive overview of the workflow for getting signed access agreements with data providers.

1 Introduction

We have in Europe a multitude of diverse medical data sources covering a wide range of domains, which holds an incredible potential for advancing research. These data sources are rich to drive research, innovation, and decision-making. Yet, accessibility to these rich repositories rises critical ethical, legal and privacy concerns. These challenges necessitate a responsible approach starting from identifying suitable candidates to negotiating and getting signed access agreements with the data owner.

In parallel to these rich medical data pools, the concept of Digital Twins has emerged as a revolutionary technology. Digital Twin offers an endless source of data also capturing patient details and providing digital versions consisting of personal health data together with stat-of-the-art digital twin models. Nowadays with the recent advancement of technologies, these digital twins can be extremely powerful in creating patient specific/stratified insights into the disease etiologies, create in-silico trials representing different sub-groups of the population. While medical data pools provide a large source of real-world patients, digital twins create virtual versions of individuals to simulate and fill-in the gaps where data is lacking. Together with the European medical data pools, the digital twin technology unlocks the ability to tailor solutions to design datasets with endless possibilities of criteria.

REALM, an ambitious initiative, aspires to bring together resources to create a substantial data repository. REALM envision to board these resources for evaluation of medical software/devices through REALM. In this context, a meticulous investigation has been conducted to assess relevant data sources required for a good evaluation of software/medical devices. In an era marked by the confluence of medical science and cutting-edge medical software/devices the necessity of evaluating application targeting the EU market is now a necessity.

The project includes five demonstrators as use case, each employing different types of data. Special emphasis is placed on investigating further demonstrators' data requirements and match them with relevant data pools. This meticulous process involves highlighting the specific needs of each use case and identifying the most suitable data sources within the vast European medical data landscape.

Finding matching data pools has proven to be a pivotal step in our journey. By strategically aligning the requirements of each demonstrator with the appropriate data repositories, we have been able to design and draft data access agreements. Final version of the agreements should take into account various factors, including the nature of the data (e.g., patient records, clinical trials, genomic data), the access type, and compliance with stringent ethical and legal standards.

2 Data Sources

2.1 Data types

Following indications given in the DMP (see DMP section 1.2, July 2023) a particular attention was given to the evaluation of the data types required for the demonstrators (Table 2). In the evaluation process for the consideration of the data pools key data types (Table 1) were identified. Mapping between data pool and demonstrators is developed further in section 5.3.1.

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), Retina Scans, ...
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 1: An overview of data types, formats, and elements in REALM

Notice that Demographics and EHR are a key data sources required by all demonstrators. EHR offer comprehensive insights into a patient's medical history, treatment, medications, and lab results. Integrated with demographics data they provide the foundation to personalized care, by optimizing treatments, track health trends and enhance patient outcome. The following table (Table 2) emphasis how demonstrators heavily rely on demographics and EHR data.

Demonstrators	Data Types	Formats	Elements
Dune AI	Demographics & EHR	Structured data	Age, sex, disease, and other clinical information
	CT Scans	Image data	Medical images of patients' lungs used by the DuneAI software for tumor segmentation in DICOM (Digital Imaging and Communications in Medicine) format.
	SUS evaluations	Structured 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
PGX2P	Demographics & EHR	Structured data	Sex, age, ethnicity, diagnosis, and medication data
	Genomics	Sequencing data	Genotyping, Whole Genome Sequencing

	Genotyping	Sequencing data	Collect from a small cohort for validating the application.
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
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.
	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
COPowereD	Demographics & EHR	Structured data	Age, sex, disease, symptoms, and other clinical information
	Lung monitoring	Time-series data	Vital signs

Table 2: Descriptions of datasets from five demonstrators in REALM

2.2 European Medical Data Pools

European Medical Data Pools represent a pivotal convergence of healthcare information across the continent, fostering an environment ripe for innovation and advancement in medical data science and AI. These expansive repositories house a myriad of patient data, clinical insights, and healthcare outcomes, serving as an invaluable asset for research and development. Mapping of different types of Medical Data Pools is an ongoing work which we will continue throughout the project to enrich the REALM data repository offerings.

2.2.1 Biobanks & EU Data Repositories

In our research of potentials data sources for REALM, we created a list (see sample in Table 3 and full list in Supplement 11) of European data sources/biobanks. It reflects the importance of assessing medical software with the highest quality and impact. Biobanks are evaluated under the following criteria's:

1. **High ethical and scientific standards**: the repositories provide high standard medical data (or in the process of standardisation).
2. **Rich and diverse data repositories:** the need of a broad range of data and data type has been highlighted by the demonstrators. This need will increase with the coming medical application that REALM is aiming to evaluate.
3. **Accessibility**: access to the repositories or through an intermediary can be negotiated and granted following terms of a Data Access Agreement.
4. **Wide data types**: a rich range of data type is necessary for REALM to succeed at evaluating various medical software.
5. **Geographical reasons and diversity**: for this deliverable a particular attention was brought to European Biobanks and European initiatives. American and Asia-Pacific data sources are still under consideration and will be consider as well in the coming months.

Data Pool	Short Description	Type of Data Available	Data Format
Estonian biobank	National biobank in Estonia containing health-related data	Genomic data, medical records, lifestyle information, etc.	Various (e.g., VCF, EHR formats, questionnaire data)
FINNGEN	Finnish Genome Center's platform for genomic research	Genomic and health data from Finnish population	Genomic data (e.g., VCF, PLINK formats)
UK Biobank	Large-scale biomedical database in the United Kingdom	Genetic, clinical, and lifestyle data of UK participants	Structured data (e.g., CSV, XML, imaging data)
Lifelines	Dutch population-based cohort study and biobank	Health data, questionnaires, biological samples, genetics	Various (e.g., structured data, biological samples)
Health RI	National infrastructure supporting health research in NL	Various health data sources, research tools, and platforms	Depends on the specific data source[i]

Table 3: Sample of data pools list. Full list available in Supplement and provides more information (URL, Country…) as well as contact progress fields.

2.2.2 Data Spaces & Initiatives

In Europe, the emergence of comprehensive data spaces and initiatives such as OpenAIRE, the European Health Data Space (EHDS) and others signify a paradigm shift towards a more interconnected and accessible digital landscape. OpenAIRE serves as a catalyst for open sharing, making a myriad of research outcomes accessible to a broader public, thus fostering a collaborative ecosystem for knowledge exchange and innovation. On the other hand, the EHDS is at the forefront of revolutionizing healthcare, aiming to securely share health data across EU Member States, thereby improving healthcare outcomes, research, and policymaking. Together, these initiatives underscore Europe's commitment to leveraging data as a catalyst for innovation, ensuring that data-driven advancements are inclusive, secure, and aligned with the collective aspirations of the European community.

2.2.2.1 Open Access Infrastructure for Research in Europe (OpenAIRE)

OpenAIRE stands as a pivotal European initiative with a foundational goal of fostering open sharing and promoting enhanced accessibility of research outcomes and inputs to a broad spectrum of the public. It serves as a rich repository for a diverse array of medical research studies, often accompanied by associated datasets and results spanning a multitude of data types. The accessibility to such datasets is invaluable, offering a rich substrate for assessing medical applications and conducting comprehensive evaluations. While OpenAIRE may not unequivocally serve as a primary data source, it unequivocally holds substantial value for REALM, proffering specific data and materials intrinsically linked to publications, thereby enhancing the depth and breadth of our research endeavors.

2.2.2.2 European Health Data Space (EHDS)

The European commission is providing funding for project facilitating data initiatives to harness the potential of data-driven innovation and facilitate cross-border collaboration. In this regard, the European Health Data Space (EHDS) is working to unlock more data and improve interoperability while

maintaining control and transparency for individuals. EHDS introduced the concept of a “data space” to centralize health data in Europe. The ultimate goal of the initiatives is to provide support for organizations seeking to connect with health data. Numerous organizations are collaborating in this effort in order to achieve a competitive, independent and sustainable European Union on the health data market.

EHDS remains an ongoing project that is not expected to be effective before a few years. However, connecting REALM with the coming European data spaces is an important set-up in achieving data interoperability, collaborative innovation and strengthen the data network. To follow-up on the EHDS initiatives, VITO committed on attending workshops, conference and seminars that gather data initiatives and actors in Europe (MEDVIA, OHDSI Europe, VITO, Imec etc…).

2.2.2.3 OMOP Federated Networks

Validation of AI requires a large and wide amount of medical data. We identified in the previous sections the various data type required for our demonstrators. In fact, much more data type and realistic data needs to be taken into account for the validation process. In this context, collaborating with OMOP (Observational Medical Outcomes Partnership) Federated Networks emerges as an incredible opportunity, offering a huge amount of medical data and metadata. This partnership stands as a chance for validating AI software with real-world, diverse, and comprehensive healthcare insights. During our investigations we have established a list of Federated Networks (Table 4 and full list in Supplement) we are considering partnering with.

OMOP Federated Networks represent a consortium of healthcare organizations, research institutions, and data custodians, united by a common goal: harness the power of vast and diverse medical data repositories. These networks are pooling together electronic health records (EHRs), claims data, patient registries, and other clinical data sources. This accumulation creates a rich and heterogeneous sample of healthcare information, capturing rich data beyond anything that an individual dataset could never achieve alone.

The value of partnering with OMOP Federated Networks becomes evident in the context of AI software validation. The effectiveness of AI algorithms in real-world medical scenarios hinges on their exposure to a comprehensive spectrum of patient demographics, disease profiles, treatments, and outcomes. OMOP's federated approach ensures that these AI models encounter a representative patient population.

OMOP Federated Networks offer not only data but also metadata which is critical for making prediction. OMOP's standardized vocabularies, ontologies, and data structures provide a coherent framework for interpreting and utilizing the data. OMOP and OHDSI tooling is also inspiring for the whole data life cycle from generating synthetic data (see WP4.4) to generating relevant cohorts (see Athena OHDSI) for assessing various AI models.

In a landscape characterized by data privacy concerns, OMOP's federated networks ensure data security and sovereignty. The data stay local, and each participating entity retains control over its data, sharing only aggregated and de-identified information. This allows to preserve patient confidentiality, and ethical standards.

OMOP Federated Network	Description	Number of data partners
EHDEN	A major initiative standardizing health data and enabling large-scale analytics using the OMOP common data model. Collaborative network with data partners across European countries.	187 Data Partners 29 countries
DARWIN	DARWIN EU collaborates with data partners who help generate real-world evidence that can be used in scientific evaluations and regulatory decision-making.	10 Data Partners
EURACAN	Collaborative initiative aimed at improving the diagnosis, treatment, and research of rare adult solid cancers across Europe.	163 Centres
HONEUR	HONEUR (Haematological Outcomes Network in Europe) is an initiative aimed at enhancing the understanding and treatment of haematological malignancies across Europe. It seeks to harness the potential of Real-World Data (RWD) to drive advancements in patient-centered healthcare models	20 Data Partners
PIONEER	PIONEER aims to transform the field of prostate cancer care with particular focus on improving prostate-cancer related outcomes, health system efficiency and the quality of health and social care across Europe.	39 partners
OPTIMA	Implementation of a vast federated and centralised network of European data providers to help answer the highest priority research questions in prostate, breast and lung cancer, especially where the existing evidence underpinning clinical practice guidelines is weak or lacking.	95 data sources (23 ready in platform)

Table 4: List of European Federated Networks. Full table is available as Supplement and provides extra fields and descriptions as well as contact status with representatives.

Because of the heavy reliance of our demonstrators on demographics and EHRs, we believe that OMOP federated networks are an essential resource as they focus on sharing structured and standardized EHR data across partners. Most medical application (including the demonstrators) rely on structured data, and OMOP is the most widespread data format for providing structured EHRs across EU. Moreover, OMOP has been enhanced and maintained with specific extensions that also fit with our demonstrators' requirements (Table 5). These extensions are getting more and more popular within the research community by including other medical standards and data type.

Extension Name	Description	Data Type	Publication	Demonstrator
G-CDM	An extension for representing genomic data within the OMOP Common Data Model, facilitating the integration of genetic information into healthcare research.	Genomic Data	(Shin et al., 2019)	PGX2P
R-CDM	Designed to handle DICOM (Digital Imaging and Communications in Medicine) files, which are commonly used in radiology and medical imaging. It allows the integration of imaging data into OMOP-based research.	Medical Imaging Data	(Park et al., 2022)	COPowereD, Dune AI, ASCOPD
Oncology Extension	An extension tailored to handle oncology-related data, including cancer diagnoses, treatments, and outcomes, facilitating cancer research within the OMOP framework.	Oncology	(Belenkaya et al., 2021)	Dune AI

Table 5: summary of widely used OMOP extensions mapped to the demonstrators.

2.3 Synthetic Data

The inclusion of synthetic data in the REALM project is crucial for providing high quality evaluations. Two main components will be incorporated in this regard: REALM synthetic data generator and DT. The combination of both approaches will cover data challenges and provide a solid base for the REALM project.

2.3.1 In-house Synthetic Data Generator (SDG)

The main aim of using synthetic data in REALM is to validate AI as a medical software. Thus, synthetic data is to be used in regulatory 'stress tests' before AI algorithms can be used by physicians and patients. For example, by creating synthetic data that simulate variable lighting or camera distortions, or that imitate data collected from devices from different vendors or from different imaging modalities, synthetic data can be used to test against model deuteriations in the face of domain shifts. Moreover, by generating synthetic data from under-represented cases, we can better evaluate AI on subpopulation level rather than on the overall level, thus mitigating the problem of hidden stratification. Thus, the aim of this component is to generate synthetic data to better evaluate AI as medical software, without privacy concerns. The quality of the generated synthetic data will be evaluated as well.

Different models (GANs, diffusion models, statistical) will be developed to generate synthetic data from the RWD in the data repository (T4.4). Different data modalities and medical applications will be taken into consideration, and the most suitable model for each application/modality will be used accordingly. We will not only generate the synthetic data, but also evaluate it. The generated synthetic data will be evaluated for its quality in terms of utility, privacy, and fidelity. Progress on the implementation of the synthetic data will be reported in D4.6, M33 (Multimodal Synthetic Data Generator).

2.3.2 Digital Twin (DT) Initiatives

European initiatives are emerging aiming to provide accurate data used for development and validation purposes. A human virtual twin can significantly solve gaps where data is missing by creating realistic data that involves generating synthetic data to replicate an individual's physiological, genetic and medical characteristics. A digital twin for medical data is an innovative concept that leverages the principles of digital twin technology to create a virtual version or representation of a patient's health and medical history for a given disease. This digital representation can include a wide range of data, such as electronic health records (EHRs), medical imaging, genomics information, vital signs, and more. The goal of a medical data digital twin is to provide a comprehensive and real-time view of a patient's health status, enabling healthcare providers to make more informed decisions and improve patient care. Here are some key aspects of digital twins for medical data:

In this regard, the Ecosystem for Digital Twins in Healthcare in Europe (EDITH) is aiming to create the first European human twin and pioneering advancements in synthetic data generation and in silico approaches (i.e organ simulation). On the other hand, The Virtual Physiological Human Institute (VPHi) through its collaborative effort, is coordinating the development of in silico models that simulate physiological process targeting aiding to the understanding of diseases and development of personalized treatments. VPHi effort aims to address challenges such as awareness and resource access that currently are critical point for the development of the European Digital Twin.

3 Access Type

Access to diverse and sensitive data sources is essential for REALM and improving evaluation quality. However, designing access agreements in a project evolving multiple data pools is complex. This section dives into the three key data access strategies: Direct Access, Indirect Access and Hybrid Access as they solve the challenges when forging partnerships with various medical data sources.

3.1 Direct Access

Direct access involves authorized individuals or entities directly interacting with primary data source. Direct access is commonly used by healthcare professionals, researchers, and authorized personnel and allow them to retrieve, analyse, or utilize sensitive data. These parties have the necessary permissions and credentials to query and access the data directly from the source. In the healthcare context, this often means access to Electronic Health Records (EHRs), patient registries, or clinical trial data. Data is used for different purposes such as patient care, medical research, and administrative tasks. It enables real-time and personalized interaction with the data.

Direct access ensures real-time data access which can be crucial for making informed medical decisions. At the moment, evaluation of the demonstrators does not require real-time access but it's definitely something that will need to be taken into account in the near future (i.e evaluation of real-time medical devices). Finally direct access allows to perform analysis on granular data without intermediaries, reducing the latency to access information and allow faster decision-making.

Direct access can also include access to biobanks and give the ability to directly obtain and utilize biological dataset. Biobanks are specialized repositories that collect and store collections of biological specimens such as blood, tissue, DNA and associated data, imaging data for research and validation purposes.

Direct access also presents significant challenges such as privacy concerns and legal and ethical issues. There are indeed strict data protection measures to protect patient information and comply with all regulatory authorities. Finally direct access does not always ensure well integrated data. As different systems and sites may use diverse formats and standards, harmonizing data from various sources can be complex. Moreover, data obtained indirectly may lack of granularity. Missing information can limit the scope of the analysis to answer some research questions or defining inclusion, exclusion criteria.

3.2 Indirect Access

Indirect access refers to obtaining information from the medical data pool through predefined reports, summaries, or aggregated datasets. Authorized parties access data that has been anonymized or de-identified to protect patient privacy. Instead of directly querying the data pool, they receive structured data outputs that provide insights while maintaining confidentiality. Indirect access is suitable for scenarios where privacy and security concerns are paramount, and it ensures compliance with data protection regulations. Thus, indirect access goes through an intermediary such as a coordination centre, a research consortium, or a federated network.

The intermediary establishes and enforces guidelines and standards for data access and sharing. It defines the practices for handling the data ensuring compliance with legal and regulatory requirements while maintaining privacy and security of sensitive information. The coordination centre also

determines who can access specific datasets and under what conditions and purposes. Formal access must be requested following agreements with the intermediary. Finally, the request is evaluated against the established guidelines before granting access.

3.3 Hybrid Access

Hybrid access combines element of both direct and indirect access strategies. It allows the integration of data from multiple sources about the same individual/population in different location. Hybrid access can happen when some part of the dataset can be directly accessible, but the other part remains on a specific platform with regulated and restricted access (i.e UK Biobank genomics or proteomics data on DNAnexus).

Hybrid data access requires ID matching for each data subject within the REALM infrastructure facilitating the access to evaluated software to reach the required data types across different resources and should be approached with care. The conditions in-which a hybrid access should be foreseen will be critically evaluated.

Thus, hybrid access can include access to Digital Twin synthetic data. Rather than requesting patient data (direct access), an intermediary (that host the Digital Twin models) can be requested to generate synthetic data. These synthetic datasets could be created by models based on patterns and structures learned from Real-World Data (RWD) during their training. Hence, access to models that generate synthetic data does not readily fall into the categories of direct/indirect or hybrid access but can be utilized by any of the three access options, depending on the data provider, data sensitivity and the requirement of access to real patient data sources for generating it.

4 Contributions & Link with other Tasks

The development of this deliverable has been a collaborative effort that reflects our commitment and strong partnerships with partners and people involved in various tasks within and outside the WP4. Many people contributed to the effort, this deliverable try to summarise all input and how it was integrated to implement data access agreements.

Throughout the progress on this deliverable, we have actively engaged with the task partners (VITO, TRAQBEAT, UM) to ensure that this deliverable aligns with the objectives. The continuous effort of the partners and their input was crucial to the success of this deliverable. In this regard, VITO was orchestrating the effort and organizing recurrent meetings and update to gather valuable input and expertise from the partners involved.

UM is the main contributor to writing the Data Access Agreements (see section 10.2). Strong collaboration with UM and especially the legal department (UM, T2.3) was necessary to construct the first draft of this document. VITO initially worked on the template of the Access Agreements to list required items and missing points were finally included in the document. Moreover, the effort on this deliverable is mainly based on the initial Data Management Plan (UM, D4.2, July 2023).

The contribution from demonstrators was also a key point for succeeding this deliverable. Especially input from TRAQBEAT (ASCOPD) and VITO (PGx2P) was essential to define requirements and relevant data pools and initiate data pools investigation.

Regarding the recent progress on synthetic data generation and agreements with the Digital Twin, commitment of the VPHi institute was clearly established and guarantee their contribution to implement access agreements and their support throughout the project. Finally, collaboration with UM and T4.4 (Synthetic clinical data & digital twin) was crucial for establishing objectives and claims in using the DT technology to leverage our synthetic data potential.

5 Agreement with European Data Pools

5.1 Objective

In this section, we further develop the effort to establish access agreements with EU data pools. Because patient population and medical practice vary widely across different healthcare institutions, it is difficult to predict how software will perform in real-world conditions. Access to relevant real medical data is crucial for the evaluation of medical software. It must undergo rigorous testing using relevant data to ensure efficacy, accuracy, safety, and reliability. This will be assessed by REALM to indicate to the medical institute whether the software has the desired performance characteristics. The diverse nature of medical data reflects the complexity of clinical scenarios, patient demographics etc. making it essential to validate software under these divers' conditions. Only real medical data can give insights on how the software will perform in the complex landscape of healthcare.

The agreement aims to facilitate access to valuable healthcare data for testing and evaluation of medical software. This section excludes potential agreements and discussion with Federated Networks and Data Spaces initiatives. As these data sources are protected and regulated by an intermediary, the implementation of access agreements is a totally different approach that needs to consider their own rules and regulations. This section focuses on direct agreements with the data owner.

5.2 Background

Access to relevant, real data is crucial for evaluation of healthcare applications. We are aware that a medical software does not always perform as expected in practice. Access to these data sources can enrich the quality of the evaluation and ensure that the medical software can be used in the real world.

5.3 Approach

1. **Identify and evaluate Data Pools**: first step in the agreement journey is to identify suitable data pools that align with our objectives (evaluation of the demonstrators). Further investigation might be required for assessing the quality of the required standards.
2. **Reach out to Data Pools**: several discussions with the data owner needs to be conducted to clarify the aim and scope of the project. Conversations are essential to clarify the project's objectives, establish trust, and ensure that both parties are aligning goals and ethical considerations. A REALM data partner brochure (see Annex 10) has been created to quickly highlight the key points of the collaboration.
3. **Discuss terms of the access agreements:** effective access agreements can be implemented following the discussions and the terms of the data owner. We have implemented a first general draft of the access agreement (Annex 10.2)
4. **Sign access agreements**: both parties acknowledge the terms defined in the legal document.

To approach data pools candidate, we also created a brochure (see Annexes 10) to advertise in a one-page document the project scope and aim and how and why they should contribute/collaborate to the project.

5.3.1 Demonstrators' contribution and requirements

In order to provide the best data to evaluate our demonstrators, we request that partners provide information about data sources (biobanks or synthetic data) used by the demonstrators internally for development and validation of the software. The section below describes the demonstrator's data requirements. A Template spreadsheet has been created and forwarded to the demonstrator. This template is requesting the demonstrator's contribution to indicate data sources they could recommend to REALM based on their experience.

Demonstrator	Leader	Person
DuneAI	UM	Philippe Lambin
PGx2P	VITO	Simon Denil
STAR	ULIEGE	Thomas DESAIVE
ASCOPD	TRAQBEAT	Vangelis Sakkalis
COPowereD	COMUNICARE	Alfred Attipoe

Table 6: Demonstrator list and statute progress on identifying demonstrators' data sources.

Initial identification of requirements from the demonstrators have already been identified in the DMP (D4.2), thus we are now considering various data sources they are using for REALM. Addressing the demonstrator requirements is crucial for providing a good evaluation. However, we are aware that REALM will need to rearrange and combine multiple sources and synthetic data to avoid the bias of evaluating medical software with the same inputs that were used for development.

Finally, we have mapped below the demonstrators with our findings following their requirements in data type (identify in section 2.1). Data requirements for demonstrators (T6.1) will be investigated further (Task start in M20) and will provide the detailed mapping with the data pools available.

5.3.1.1 Dune AI – Maastricht University

Dune AI (REALM Proposal Section 1.3.9 Use Case 1 Page 18) relies on existing patients' CT Scan data and collection of new patients' demographics, CT scans, System Usability Scale (SUS) evaluation data, and qualitative feedback from clinicians.

At the moment, data sources for testing and development of Dune AI are still under investigation at UM. However, as Dune AI heavily relies on lung imaging and cancer data, multiple data pools have been identified already (see 11). On top of this, we have already described in 2.2.2.3 how an Oncology Oriented Federated Network (i.e., HONEUR) could be highly valuable for such a demonstrator.

5.3.1.2 Pharmacogenomics passport (PGX2P) – VITO

PG2PX (REALM Proposal Section 1.3.9 Use Case 2 Page 19) initially planned to heavily rely on UK Biobank for development purpose and is still considering that option.

PGx2P demonstrator is still in the process of evaluating data sources as the project is in the early stage of the development. A few data sources candidates have been identified (Table 7) for development purposes. Data sources and requirements will be updated (T4.2) following the feedback from the experts in a scheduled PGx2P demo beginning of October 2023.

Data Pool name	Short Description	Data Format
GeT-RM	Multiple datasets on cell lines characterized by WGS, targeted sequencing and other molecular assays to determine alleles relevant to pharmacogenetics. Subset of the raw data is publicly available through European nucleotide archive.	CSV, fastq/CRAM
Jessa Patients Dataset	Planned dataset to be collected from a cohort of patients in a Belgian hospital in the context of pharmacogenomics study.	CSV, fastq/CRAM

Table 7: Data pools contribution from PGX2P (VITO)

5.3.1.3 STAR – Liege University

STAR (REALM Proposal Section 1.3.9 Use Case 3 Page 20) relies on clinical patients' data including their demographic and clinical data focusing on glucose and patient response to insulin and questionnaire data for usability assessment.

Data Pool name	Short Description	Data Format
D1NAMO	A multi-modal dataset for research on non-invasive type 1 diabetes management (Dubosson et al., 2018)	--

Table 8: Data pool candidate for STAR.

Several options are considered for STAR which require specific data types such as diabetic and health patient data collected from wearables (Table 8).

5.3.1.4 ASCOPD – TRAQBEAT

ASCOPD (REALM Proposal Section 1.3.9 Use Case 4 Page 21) relies on existing patients' data including demographic and clinical data from their admission and discharge events and generates new heart-related data using TRAQBEAT sensors.

TRAQBEAT development dataset mainly comes from a Greek hospital. Contacts within the institute have been provided by TRAQBEAT and are sufficient to get in touch with the data department. The data sources are currently under evaluation and may be reached following our approach (see section 5.3) after discussion of terms of our Access Agreements (developed in Annexe 10.2).

Data Pool name	Short Description	Data Format
Ambulatory Health Care Data	Data on the utilization and provision of ambulatory care services in hospital emergencies. Identified adults with asthma or COPD exacerbation.	CSV
COPD Patients Dataset	The dataset is a cohort of patients, who had been invited to take part in a rehabilitation program to help manage their COPD.	CSV
In Hospital Mortality Prediction	Predictors of in-hospital mortality for intensive care units (ICU)-admitted HF and COPD diagnosed patients	CSV
Lancet dataset:	Data used from the Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE, a three-year observational study, 1,803 patients, 2,117 exacerbations) study – (*Recalibrating and externally validating the Acute COPD exacerbation prediction tool*)	--
Nature scientific reports:	The medical records of 410 hospitalized AECOPD patients are collected. (Peng et al., 2020)	--

Table 9: Data pools contribution from ASCOPD (TRAQBEAT)

5.3.1.5 COPowereD – COMUNICARE

COPowereD (REALM Proposal Section 1.3.9 Use Case 5 Page 22) relies on observational data from patients with chronic obstructive pulmonary diseases from hospitals.

At this moment, dataset for COPD can be found following (Swaminathan et al., 2017) study.

Data Pool name	Short Description	Data Format
COPD	Dataset. Training and validation dataset of COPD patients -- (Swaminathan et al., 2017)	CSV

Table 10: Data pools contribution from CoPowereD (COMMUNICARE)

5.3.2 Implementing Agreements

The contribution of the legal department of Maastricht University was crucial to implement a professional agreement document. The template (Annexe 10.2) of the access agreement will be used to negotiate the terms of the agreements between the RECIPENT (REALM instance) and the PROVIDER (data pool).

The access agreement consists of three parts, each part essential and legally binding. The main body of the access agreement outlines the terms, conditions and parties' obligations regarding the access to data. In addition, two legally binding annexes are added to the access agreement. Annex I require a description of the data, an overview of the methods of transfer and storage, and an overview of the allowed processors. Annex II ('The Research Plan') will outline the specific purpose(s) for which the data may be used.

Concluding on an access agreement between the REALM instance and the data pool is an essential requirement. The exact terms and conditions between the recipient (REALM) and the provider (data pool) must be discussed further. Annex 10.2 will serve as the basis for these further discussions.

5.4 Confidentiality and Data Protection

This section lists crucial confidentiality and data protection aspects that have been developed further in the agreement template (see Annexe 10.2).

5.4.1 Compliance with GDPR and Applicable Data Protection Law

Compliance with the General Data Protection Regulation (GDPR) and applicable national data protection laws are essential to the data access agreements. To comply with GDPR the access agreements shall include the requirements below:

- Fulfilment of the data protection principles as required by Article 5 GDPR.
- Guarantee of appropriate technical and organizational measures to meet requirements for data controllers (GDPR article 32.1).
- Ensure parties cooperation in case of data breach (GDPR articles 33, 34)
- Ensure that any data processors are instructed to process personal data in accordance with the requirements stated in the GDPR
- The data shared by the data owner shall constitute pseudonymized personal health data under the GDPR. Anyone with access to the data shall agree to not conduct any procedure with the intention of determining or revealing the identity of the patients.

5.4.2 Technical and organizational measures

The data access agreements ensure that both parties (PROVIDER and RECIPENT) will provide appropriate technical and organizational measures. The data access agreement stipulates that data shall be provided by the PROVIDER in a secure manner. The RECIPIENT shall implement appropriate security measures to guarantee the safe storage and usage of the data. The RECIPENT also consents to

its statutory notification obligations in case of data breach. Any data leak must be treated following the 'personal data breach' terms (GDPR, articles 33 and 34). Training and awareness for data usage and security measures should be provided to bodies and partners. We are considering including this training program within REALM Academy (T7.4), future discussion with T7.4 will be scheduled to set the content and format of the training program.

The temporal limitation is another key element of the data access agreement. Data owners may usually use expiration date for data access as an extra layer to protect their data. Thus, data is accessible in a specified timeframe only as agreed upon in the access agreement, thereby minimizing the risks of unauthorized or prolonged data access. The period can only be extended by mutual agreement in writing by both Parties.

5.4.3 Data processor, data controller, and joint data controller-agreements

The GDPR defines the concepts of 'data controller' (article 4 GDPR). For the purpose of this research project, the provider ('the data pool') shall be considered the data controller of the data under the GDPR up until the moment the data is provided to the recipient ('the REALM instance'). From that moment onwards, the recipient ('the REALM instance') shall be considered to be a separate data controller under the GDPR and the applicable national data protection laws for the processing of the data for the purpose of the REALM's research project.

The data controller determines the purposes for which and the means by which personal data is processed, while the data processor processes personal data only on behalf of the controller. It is also possible for multiple organisations to be a 'joint controller' if they jointly determine 'why' and 'how' personal data should be processed. Joint controllers must enter into an arrangement setting out their respective responsibilities for complying with the GDPR rules. The main aspects of the arrangement must be communicated to the individuals whose data is being processed[3].

In the REALM project, joint data controller agreements will be established between the REALM consortium and the data providers to make clear their respective responsibilities for complying with the GDPR rules and applicable national data protection legislation, as required by Article 26 GDPR Data subjects will be informed under the requirements by Article 14 GDPR. The key aspects of the arrangement will be communicated to the individuals whose data is being processed, in accordance with the GDPR.

5.4.4 Material Transfer Agreements (MTA)

A Material Transfer Agreement (MTA)[4] a legal document that governs the transfer of tangible research materials between two organizations, when the recipient intends to use it for research purposes. It's a crucial contract that clarifies the rights, responsibilities, and obligations of both the provider and the recipient with regard to the materials and any derivatives. MTAs protect both the providing and the receiving party on ownership rights, academic publication and priority (who gets to publish first), intellectual property (who owns improvements and new inventions relating to materials) and permitted use and liability.

The REALM consortium will establish material transfer agreements with data providers where needed to make sure all rights, obligations and responsibilities are clear.

[3] European Comission: *What is a data controller or a data processor?*
[4] Material Transfer Agreements: *What is an MTA?*

5.4.5 Ethics & Informed Consent

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 committee will be responsible for reviewing data access requests, ensuring that the data is only shared with parties that have a legitimate research need, and that appropriate measures are in place to protect patient privacy and the proprietary aspect. The committee would also ensure that all data use is compliant with relevant laws and regulations, as well as the consortium's data use agreements.

5.4.6 Pseudonymized data, Anonymization and Deidentification

Pseudonymized data, anonymized data, and deidentified data are all techniques used to protect the privacy of individuals. They differ in terms of how reversible the process is and how much information is retained. Pseudonymized data replace all personally identifiable information by a pseudonym (or token). Data can still be linked back to the original individual using a separate key that should be safely kept separately. In the other hand, anonymization is a stringent irreversible process that removes all identifying information.

Anonymization or deidentification are not cover by GDPR regulations. For the purpose of this project, only pseudonymized data shall be processed, as defined by Article 4 of the GDPR. As stated in the access agreement, the provider shall ensure that the data are pseudonymized before providing the data to the recipient. In this regard, the recipient acknowledges to verify the effectiveness of the pseudonymization. The Recipient shall not attempt to reverse engineer, de-identify, or re-identify the Data to identify individuals or entities.

5.5 Data Storage & Access

5.5.1 REALM Federated Network & Infrastructure

Sensitive and pseudonymized data can be stored in public research data repositories such as Physionet or Dryad with restricted data access control. This in-principle can give flexibility to application developers on how to provide their "test" data to the REALM infrastructure in a standardized, structured and secure manner. We are exploring this in REALM and Data storage and infrastructure will be elaborated further as described in the T5.2 (Data Fusion and System Integration). Moreover, the REALM governance and technological architecture will be established in T3.1. We already got in touch with the partners to align on dependencies with T3.1 and T5.2. In this regard, input from partners involve in the deliverables will be crucial to ensure compliance with ongoing agreements with the data pools.

Figure 110.2

At this stage in the REALM Project lifecycle, the design of the Federated REALM Architecture (Task 3.1) aims to also accommodate the various data access and storage challenges. Taking in consideration the different data access preconditions, limitation and the available legislation for data handing and curation, such as the GDPR, the REALM architecture accommodates federated data access, along with data localisation limitation, by supporting federated platform instances. In this scheme, each REALM instance is able to include any data sources following the data access agreements template (10.2), from, various locations in a federated fashion (Figure 1). This strategy can prevent some data access limitations (i.e., country restrictions) and provide tailored access agreements following the template we developed for this deliverable. This scheme takes in account the different requirement of the variety of stakeholders, in respect also to data owners, that will be involved or use the REALM platform. The data transferability, curation, processing and serving will be considered in the design process. The complete Federated REALM Architecture will be developed and presented in the subsequent deliverable “The Federated REALM Architecture” D3.1 by M28.

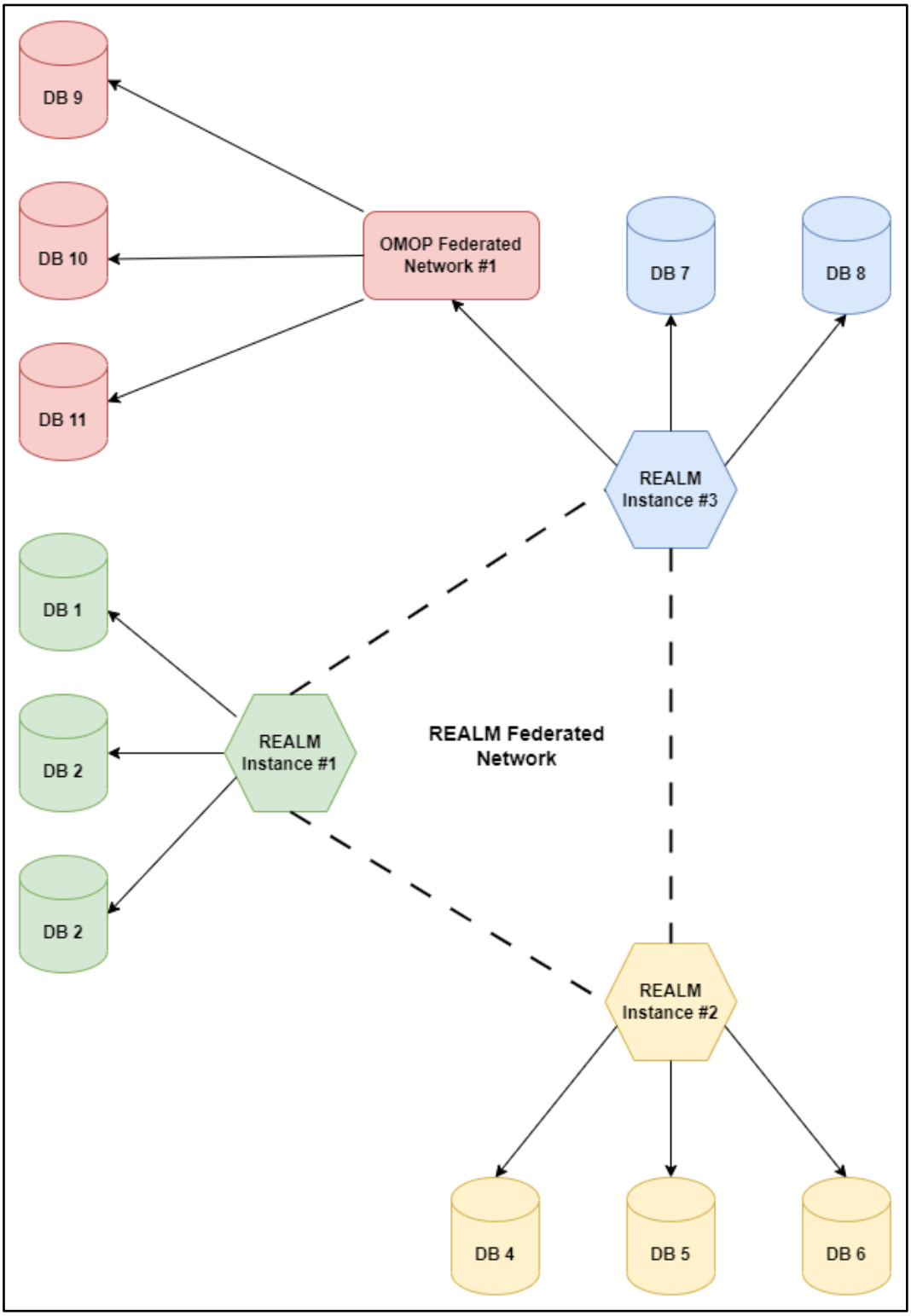

Figure 1: REALM Federated Network: current vision of the realm federated case. This diagram is given as an indicative concept which is still under development. Several instance of REALM (nodes) could be connected together with access to common and different data source. REALM Instance #3 also show case that some of the nodes could be connected to external data spaces or Federated Networks (i.e OMOP federated networks).

5.5.2 Blockchain Smart Contracts

One important note is the planned implementation of light weight Blockchain Smart Contracts (BSC). "A Blockchain Smart Contract is a computer program that (directly and automatically) determines the transfer of digital assets/data between processes (or parties) under predetermined and agreed conditions. The use of smart contract ensures the 1) storing of the metadata regarding the data resources used, 2) versions of executed analyses and 3) links to the resulting analyses. In REALM we are investigating the level which lightweight BSCs can be used. Our primary goal is to capture the provenance of the process followed under the hood of the REALM infrastructure rather than actually executing the data exchange with external data sources, hence is not extensively covered in this deliverable. However, we are also exploring the potential of BSCs as a candidate process to access external data sources.

This highly depends on the willingness & capability of external data providers to implement them at their side for data exchange. If included, they will be developed further in the next version of the DMP. Furthermore, BSCs can also be used to identify "similar" data needs across various instances that can be accessed (if the data provider agrees to share the data for the same intent of the data requestor, or in other words there is a match between the data provider and requestor ends). This type of access request comes also with transparency in the REALM architecture and across REALM nodes about which data were considered for model testing, validation, and certification.

5.6 Result

Thanks to the collaboration of the demonstrators and our continuous effort on identifying data sources, we have initiated the data mapping targeting the demonstrators' requirements. Thanks to this achievement, we managed to implement a data access agreement template enriched by the current progress on related tasks (data storage, infrastructure, user requirements, research project…).

We are now heading towards the finalisation of robust access agreements that aligns our objectives while maintaining the highest standards of data protection and compliance with the EU regulations. The terms and conditions of REALM and the data pools still need further discussion. Thus, the draft of the access agreements we implemented serves as a template, as a basis of discussion. The access agreements have been designed to handle future requirements we have identified in this deliverable. Agreements will continuously be updated following the project evolution, regulations, and data pools requirements. The agreement template will be reviewed further later this year (October 2023) to follow up on the agreements with UM and VITO legal experts. At a later stage, Annex II (research plan which includes the purpose of the processing) of the access agreement will be updated as well according to the context of the data pool.

6 Agreement with Digital Twin (DT) Initiatives

6.1 Objective

This section documents the effort to develop agreement with the parties involved in the development of Digital Twins. We are here describing the current progress on the EDITH project and the contribution of the VPH institute developing Digital Twins (DTs). The contribution of all parties is an essential data source for the REALM project. Agreements with representatives of the DTs shall set the terms and condition for accessing the Digital Twin capability of generating synthetic data for REALM.

6.2 Background

Synthetic data generation is a key component of the project (see T4.4: Create multimodal synthetic data) to create relevant dataset for development, testing and evaluation of medical/healthcare software. It has been identified that a collaboration with the institute developing the Digital Twin could significantly benefit the REALM Project. A digital twin for medical data is an innovative technology that creates a virtual replica or representation of a patient's health and medical history. This synthetic data will play a crucial role in enhancing the depth and coverage of the REALM evaluations while ensuring privacy and security of sensitive medical information. In this regards the EDITH[5] project to implement the first European virtual human ecosystem and provide models (twins) that generate various synthetic data types.

6.3 Approach

We have listed below the key action items leading to the implementation of robust access agreements:

1. **Identify key actors in Digital Twin development**: a robust communication has been established with VPHi institute which coordinate EDITH development.
2. **Define commitment of the partners**: outline the commitment of each actor in the success of getting access agreements with the DT (Annexe 10.3).
3. **Implement agreements**: Implement together with the support of the VPHi institute the access agreement to the DT technology to generate on-demand synthetic data for REALM use cases. The access type has to be identify following the discussion with the VPHi institute and the development stage of EDITH.
4. **Signed access agreements**: finally, agreements should be signed between both parties and initiate task related to the related infrastructure and code requirements.

EDITH remains an ongoing project, a strong and recurrent communication with the VPHi institute has been established to follow-up on the project development.

6.4 Data Storage & Access

Due to the different nature of data used in the project, we consider enabling access to open-access and public research repositories such as Figshare in addition to an internal secured environment. We are investigating the ease of access and potential needs to develop this methodology. At this moment, the access type to the output of the DT is still under discussion. We have identified the most likely scenarios: (1) indirect access though the Virtual Physiological Human Institute (VPHi) via another

[5] VPH Institute - EDITH: an ecosystem for digital twins in healthcare in Europe.

intermediary that handle the security and privacy restrictions and (2) direct access to the data in a REALM environment directly hosted by REALM (or in an external public repository).

Synthetic data generated from *source* real data should be hosted with the highest security as well as its potential identifiable information about unique individuals from the *source* population. A secured cloud environment has been discussed to host DT output. Several data storage strategies are under discussion with T4.4 (synthetic data and DT), T4.2 (Data repository catalogue & management) and T5.2 (data fusion and system integration).

6.5 Result

In light of collaborative synergies with our demonstrators and persistent efforts directed towards pinpointing precise requirements, we have embarked upon an extensive data mapping exercise which the first version is reflected in this deliverable. This foundational step has been instrumental in the articulation and establishment of a comprehensive data access agreement template attached below in this deliverable. The template is a living document, reflecting the evolving advancements in related tasks such as data storage, infrastructure development, user requirements, and research project progression.

As we forge ahead, our focus is unwavering on finalizing access agreements that seamlessly aligns with our overarching objectives. Paramount to this endeavor is the commitment to uphold the highest echelons of data protection, ensuring stringent adherence to EU regulations. Anticipating the dynamic nature of project evolution, regulatory landscapes, and data pool requisites, we remain vigilant, adapting our access agreements accordingly.

A detailed review of the agreement template is scheduled for October 2023 at the REALM Project Meeting at Brussels, involving legal connoisseurs in the consortium, ensuring it remains contemporaneous and compliant.

7 Conclusion

Our explorative journey in the realms of Digital Twin technology and European Medical data pools has unearthed substantial potentials and invaluable resources. On the digital twins domain, VITO and VPHi are part of the EDITH project and we aim to utilize the results of EDITH in the REALM project. Currently, we have included a statement regarding their contribution in the form of a letter of commitment (Annexe 10.3)

Through methodical research and rigorous evaluation, we have got in touch with relevant data repositories and data spaces initiatives, aligning seamlessly with our demonstrators' prerequisites. This pivotal advancement in the identification of candidate data pools has been the cornerstone in the implementation of our versatile agreement template, allowing us to foresee and address potential challenges, including access types, restrictions, and data storage.

We are on a progressive trajectory, meticulously weaving a robust data network for REALM, fostering an optimal environment for the thorough evaluation of our demonstrators. Assured access to these data sources is fortified by our steadfast access agreements and the integration of EDITH, necessitating a collaborative ethos amongst all project stakeholders.

While the final data strategy is still a subject of deliberative discussions, primarily within WP4, we anticipate the unveiling of a comprehensive and refined REALM data approach in the DMP v2 (D.4.3, M30). This iteration will encapsulate the nuances of our ongoing dialogues with data pool proprietors and the current status of access agreements. In concert with other tasks, the refined access agreement is poised to facilitate the seamless amalgamation of medical data pools and Digital Twins within the REALM ecosystem.

Our continuous endeavors aim to synchronize synthetic data strategies with real-world data, ensuring a balanced and versatile platform that adheres to the highest standards of integrity, reliability, and compliance. This synthesis underscores our commitment to fostering innovation and advancing the frontiers of medical data science through the REALM platform.

8 Next steps

In light of the critical importance of accessing (1) EU medical data pools and (2) DT platform for a robust evaluation of medical software with REALM, we have defined the next steps as follow:

1. Finalize Access Agreement with data pools: We will keep updating the agreements with the data pools that define the terms, conditions and guidelines that govern our access to valuable resources. Update will follow coming discussion with the data owners and dependencies with other tasks such as the legal evaluation (UM T2.3), system integration (EXUS T5.2) and Synthetic data (UM T4.4).
2. Progress on Access Agreement with DT: continuous communication with VPHi and EDITH will be established to follow progress on the Digital Twin agreements.
3. Enlarge REALM data pools network: we will keep contacting and discussing with relevant data owners to arrange access agreements. We will keep the effort on investigating current and coming European data initiatives (i.e EHDS, OMOP European Federated Networks…)

These crucial next steps will ensure that the REALM project can proceed with accessing data or data generator legally and ethically responsible while ensuring highest standards of data protection and privacy.

The access agreements will undergo continuous updates and revisions throughout the project lifecycle to ensure it remains up-to-date and reflective of the evolution of the project until having a final signed agreements with all parties. Status on T4.3 and general finalized access agreements will be reported in D4.3, M30. D4.5 will be used as input for various tasks such as T4.4, 3.1, 7.4, 2.3 and 5.2. Progress and update on the agreements will be communicated to the task leaders.

9 References

Belenkaya, R., Gurley, M. J., Golozar, A., Dymshyts, D., Miller, R. T., Williams, A. E., Ratwani, S., Siapos, A., Korsik, V., Warner, J., Scott Campbell, W., Rivera, D., Banokina, T., Modina, E., Bethusamy, S., Morgan Stewart, H., Patel, M., Chen, R., Falconer, T., … Reich, C. (2021). Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *American Society of Clinical Oncology*, 12–20. https://doi.org/10.1200/CCI.20

Dubosson, F., Ranvier, J. E., Bromuri, S., Calbimonte, J. P., Ruiz, J., & Schumacher, M. (2018). The open D1NAMO dataset: A multi-modal dataset for research on non-invasive type 1 diabetes management. *Informatics in Medicine Unlocked*, *13*, 92–100. https://doi.org/10.1016/j.imu.2018.09.003

Park, C., You, S. C., Jeon, H., Jeong, C. W., Choi, J. W., & Park, R. W. (2022). Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. *Yonsei Medical Journal*, *63*, S74–S83. https://doi.org/10.3349/ymj.2022.63.S74

Peng, J., Chen, C., Zhou, M., Xie, X., Zhou, Y., & Luo, C. H. (2020). A Machine-learning Approach to Forecast Aggravation Risk in Patients with Acute Exacerbation of Chronic Obstructive Pulmonary Disease with Clinical Indicators. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-60042-1

Shin, S. J., You, S. C., Park, Y. R., Roh, J., Kim, J. H., Haam, S., Reich, C. G., Blacketer, C., Son, D. S., Oh, S., & Park, R. W. (2019). Genomic common data model for seamless interoperation of biomedical data in clinical practice: Retrospective study. *Journal of Medical Internet Research*, *21*(3). https://doi.org/10.2196/13249

Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., Kappel, G., & Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0188532

10 Annexes

10.1 REALM Brochure

REALM

Real-world-data Enabled Assessment for health regulatory decision-Making

About REALM:

REALM Vision:

REALM strives to drive forward innovative solutions that improve patient outcomes and empower healthcare practitioners to deliver the best possible care. The REALM consortium envisions a future where patients can benefit from cutting-edge medical software that is developed and evaluated with a focus on safety, efficacy, and usability, thanks to the joint efforts of our diverse stakeholders.

REALM Mission:

REALM aims to develop a collaborative framework through which regulatory authorities, software developers, healthcare professionals and policy offers can jointly create and evaluate innovative medical device software – for the direct benefit of patients and healthcare practitioners.

REALM in Numbers:

REALM is a 7.0 million euros project funded by the European Union (EU) for a 4-year duration dedicated to the project development. REALM involve 15 partners (Universities, Companies, Research institutes) from 6 different countries working together for the success of the project. REALM started in 2023 and is expecting a release beginning of 2028.

Why We Need You:

In the REALM of advancing medical care, your valuable contribution to the medical data pool is crucial. Validation of medical software require a broad a large range of data. We believe that the inclusion of your dataset could provide a particularly insightful input.

What You Can Gain from Joining Us:

Joining the REALM project will give you the opportunity to integrate your data with diverse European data sources. We can also assist you in (1) the **standardization and *fairyfication*** process of your data to medical standards (such as OMOP) and (2) anonymization and de-identification of your data. Also, this can be an opportunity for **networking** with European data partners and later join federated networks. You will have the chance to join the **REALM Academy** that will provide several trainings and workshops. Finally, joining REALM will facilitate your future requirements when you need to **evaluate AI and medical software**. Benefits of joining the project will of course be discuss during the initial meeting.

What We Expect from You:

To join the data pools of the REALM project you are expected to give your input for us to create tailored access agreement to your data. In the case of Electronic Health Record (EHR), data include both medical records & clinical events (such as measurements, observations) and patient metadata (sex, age, ethnicity ...). We will arrange a workshop to discuss and clearly define access agreements. Finally, on later stage of the project development, our developer team will reach you out arrange connection details and/or support you in the implementation of a tunnel linking your data to the REALM platform according to the access type defined in the access agreements (direct/indirect access, removal of the data from the system on termination).

How to Get Involved:

If you are interested in becoming a part of the REALM journey, you will find all the information in the section below to schedule an introduction presentation.

We will (1) introduce the project (Aim & Objectives, Partners, Team, use case, Impact, your role), (2) arrange workshop to discuss data access agreements to outline the term and conditions under REALM is granted to access to certain data. You will also find complementary information on the REALM website. Please do not hesitate to reach out if you require any further information.

Contact Information:

Frederic JUNG
Lead access agreements
Frederic.jung@vito.be
+32 (0) 456 41 64 38

Gökhan ERTAYLAN
Research lead
gokhan.ertaylan@vito.be

Thank you for considering our invitation to collaborate on REALM Project. We believe your input would be a significant asset to our project, and

Funded by the European Union

vito

tragbeat

Maastricht University

See partners full list on the REALM website...

10.2 Data Access Agreement Template

AGREEMENT ON THE ACCESS AND PROCESSING OF PERSONAL DATA

This agreement (hereinafter referred to as "Agreement") is made and entered by and between:

<…>, [CONTACT DETAILS], legally represented by the undersigned, hereinafter referred to as "PROVIDER"

and

REALM, [CONTACT DETAILS], legally represented by the undersigned, hereinafter referred to as "RECIPIENT".

PROVIDER and RECIPIENT hereinafter jointly referred to as "Parties" and individually as "Party";

WHEREAS

a. PROVIDER has generated DATA as further defined below;
b. RECIPIENT, through **<…>**, has requested PROVIDER, through **<…>,** to provide RECIPIENT with the DATA for the purpose of its RECIPIENT'S RESEARCH PLAN;
c. The purpose and means have been determined by RECIPIENT;
d. PROVIDER is willing, subject to the terms and conditions of this Agreement, to provide the DATA to RECIPIENT.

I Definitions

1. DATA: the data being transferred under this Agreement is the data that is further specified in Annex I to this Agreement, provided without directly identifying personal information. The DATA constitutes pseudonymized personal health data under the GDPR.
2. RECIPIENT'S RESEARCH PLAN: The research plan specified in Annex II to this Agreement for the purpose for which the DATA may be used.
3. EFFECTIVE DATE: The date of last signing of this Agreement.
4. INVENTION: any invention, discovery, improvement, material, signal, process, formula, knowhow, or other innovation related to or arising from the use of the DATA and/or CONFIDENTIAL INFORMATION, whether patentable or not, and obtained as a result of the performance of
   RECIPIENT'S RESEARCH PLAN.
5. CONFIDENTIAL INFORMATION: All information, know-how, grant applications, method of work, techniques, expertise of PROVIDER regarding the DATA, its characteristics and PROVIDER's research concerning the DATA, whether of a scientific, technical, engineering, operational, or economic nature, supplied to or obtained by RECIPIENT in written form, in the form of drawings or in the recording of oral conversation, or samples.
6. GDPR: the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation).
7. APPLICABLE DATA PROTECTION LAW*:* the GDPR and any additional locally applicable data protection legislation.
8. SUBJECT(S): shall mean the patient or other person from whom the DATA was obtained.

II. Terms and Conditions of this Agreement:

1. The DATA and any other information provided is made available as a service to the research community and no ownership rights in the DATA and any other information shall be obtained by RECIPIENT under this Agreement.

2.

a. DATA shall be provided by PROVIDER in a sufficiently secure manner and Parties shall handle all DATA in accordance with the APPLICABLE DATA PROTECTION LAW and shall keep such DATA confidential without any of the exclusions contained in Article 11 below.
b. With respect to the DATA, RECIPIENT shall be considered to be a separate data controller under the APPLICABLE DATA PROTECTION LAW for the processing of the DATA for RECIPIENT'S RESEARCH PLAN.
c. RECIPIENT shall implement appropriate technical and organizational measures to meet the requirements for data controllers of the APPLICABLE DATA PROTECTION LAW, hereby taking into account the nature, scope, context and purposes of Processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons to ensure a level of security appropriate to that risk, including, as appropriate, the measures referred to in Article 32(1) of the GDPR.
d. If RECIPIENT becomes aware of a personal data breach, RECIPIENT shall promptly notify PROVIDER. In such a case Parties will fully cooperate with each other to remedy the personal data breach, fulfil the statutory notification obligations timely and cure any damages. The term 'personal data breach' refers to articles 33 and 34 of GDPR.
e. In the event that SUBJECT withdraws his/her permission for the use thereof, PROVIDER shall supply RECIPIENT with sufficient information and RECIPIENT shall immediately cease all use of the relevant DATA and shall delete all copies of the relevant DATA. Upon request from PROVIDER, RECIPIENT shall confirm in writing the complete deletion of such DATA.
f. PROVIDER shall be data controller of the DATA under the GDPR up until the moment the DATA is provided to RECIPIENT.

The Parties' contact details for inquiries regarding handling and protection of DATA are as follows:

For RECIPIENT, to:

Name:
Address:

Tel: +
Fax: +
e-mail:

For PROVIDER, to:

Name:
Address: e-
mail:

3. RECIPIENT shall not carry out any procedures with the DATA, such as linking, comparison, processing, with the intention of deriving the identity of the Subject. The RECIPIENT and the RECIPIENT SCIENTIST agree that the DATA: (a) is to be used only for the academic purposes as described in RECIPIENT'S RESEARCH PLAN; *(b) will not be used for other, including commercial purposes.* Furthermore, in carrying out the RECIPIENT'S RESEARCH PLAN, RECIPIENT shall not allow third parties that are not expressly mentioned in the Annexes to access or otherwise process the DATA without prior written approval of PROVIDER. However, as an exception to the foregoing, such prior approval shall not be required for service providers in the context of the standard business operations of RECIPIENT, such as parties who supply ICT infrastructure maintenance. RECIPIENT will safeguard that any data processors who have access to the DATA are instructed by a binding agreement to process the personal data in accordance with the requirements stated in the GDPR.

4. Upon request, RECIPIENT'S SCIENTIST shall keep PROVIDER'S SCIENTIST informed of the results arising from RECIPIENT'S RESEARCH PLAN.

5. *RECIPIENT will report any INVENTIONS to PROVIDER and PROVIDER'S SCIENTIST. RECIPIENT shall promptly provide PROVIDER with a detailed written description of the INVENTION and indicate the role, if any, of any of RECIPIENT's employees in creating the INVENTION. Ownership of INVENTIONS will follow inventorship. Where ownership of any INVENTIONS vests in RECIPIENT, PROVIDER shall have a perpetual nonexclusive royalty free license to use such inventions for its internal research and teaching purposes. In the event the INVENTION is a joint INVENTION, both Parties shall make appropriate mutual arrangements concerning the protection and exploitation of such joint INVENTION. Until such agreement is effective, each Party shall be entitled to use the joint INVENTION for research purposes, but neither Party shall be entitled to exploit, disclose, license or transfer its rights in connection with the joint INVENTION.*

6. Except as provided in this Agreement, no express or implied licenses or other rights are provided to the RECIPIENT under any Intellectual Property (IP) rights of PROVIDER.

7. The DATA will be accessible by the RECIPIENT at no financial cost.

8. DATA will be accessible by the RECIPIENT in a sufficiently secure manner and in a format to be agreed upon by the RECIPIENT SCIENTIST and the PROVIDER's SCIENTIST.

9. PROVIDER warrants a) that it has verified that there is an appropriate legal ground for the provision of the DATA to RECIPIENT in accordance with the GDPR (such as Article 6 and/or 5.1 sub b GDPR) b) that there is a valid exception to the prohibition for processing personal health data (Article 9 GDPR) and c) that it shall be provided under approval from the relevant ethics committee to the extent required. Apart from this, it is

expressly understood that PROVIDER does not make any warranties regarding the DATA and specifically does not warrant or guarantee that the DATA will be accurate, be merchantable or useful for any particular purpose. PROVIDER cannot and shall not be held liable for any claims or damages by RECIPIENT or any third party, in connection with or as a result of the use of DATA by RECIPIENT. Unless and to the extent caused by PROVIDER's gross negligence or wilful misconduct, RECIPIENT undertakes to hold harmless PROVIDER at all times against all of such damages or claims.

In regard to the DATA and personal data breaches, RECIPIENT shall be responsible and liable for any damages, losses and fines resulting from its own actions or failures to adhere to the terms of this Agreement and APPLICABLE DATA PROTECTION LAW and RECIPIENT shall indemnify and hold harmless PROVIDER for any of such damages. For the purposes of this sub clause, actions or omissions of data processors contracted by RECIPIENT, shall be attributed to RECIPIENT.

10. RECIPIENT agrees in its use of the DATA to comply with all applicable international and national laws, statutes, regulations and guidelines.

11. RECIPIENT shall treat all CONFIDENTIAL INFORMATION as confidential for the duration of this Agreement including any extension thereof and thereafter for a period of five (5) years following termination or expiry of this Agreement. Excluded from this obligation of confidentiality shall be any CONFIDENTIAL INFORMATION of which the RECIPIENT can reasonably demonstrate that it (a) was previously known to RECIPIENT, or (b) is, and/or becomes, publicly available during said five (5) year period through no fault of RECIPIENT, or (c) is independently and lawfully developed by the RECIPIENT, or (d) was published or otherwise disseminated in accordance with the publication procedure set out below in article
12. However, the foregoing exceptions shall not apply to: (a) CONFIDENTIAL INFORMATION contained within more general information that may fall within one or more of the exceptions, or (b) any combination of features or items of CONFIDENTIAL INFORMATION where one or more of the relevant individual features or items (but not the combination itself) may fall within one or more of the exceptions. The obligation of confidentiality shall not apply to any disclosure required by law, provided that RECIPIENT shall notify PROVIDER of any disclosure required by law in sufficient time so that PROVIDER may contest such requirement, if PROVIDER so chooses.

12. Parties acknowledge the importance of disseminating the results of the RECIPIENT'S RESEARCH PROJECT. Therefore, RECIPIENT shall endeavor to publish or otherwise publicly disclose information, any data, results or information generated using the DATA ("Disclosure(s)"), after review by PROVIDER. The following shall apply to Disclosures: a. Authorship of any publications shall follow the principles set out in the ICMJE recommendations 'Defining the Role of Authors and Contributors' as can be found on http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-thehttp://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.htmlrole-of-authors-and-contributors.html .

b. At least thirty (30) days before RECIPIENT submits a paper or abstract for Disclosure, RECIPIENT shall provide such paper or abstract to PROVIDER, who will have thirty (30) days to review proposed manuscripts and fifteen (15) days to review proposed

abstracts to assure that its CONFIDENTIAL INFORMATION is protected. It is agreed that RECIPIENT will fully comply with any reasonable written request by PROVIDER to omit specified CONFIDENTIAL INFORMATION of PROVIDER from such paper, abstract, press release or other disclosure prior to Disclosure.

c. In every Disclosure by RECIPIENT based upon results obtained from the research through the help of the received DATA provided by PROVIDER, RECIPIENT shall appropriately acknowledge PROVIDER and PROVIDER'S SCIENTIST as contributor of the DATA.

13. This Agreement will become effective on the EFFECTIVE DATE and will terminate [...] years after the EFFECTIVE DATE, unless mutually extended in writing by both Parties. Any clauses which will be expected or intended by its nature to survive the termination or the expiration of this Agreement, shall survive the termination or the expiration of this Agreement.

    Upon expiration or termination of this Agreement, the right to use the DATA and CONFIDENTIAL INFORMATION will automatically end. However, RECIPIENT may retain one copy of the DATA solely to ensure that the DATA are findable, accessible, interoperable and reusable if any third-party academic institution requests the aforementioned, in light for replication of the RECIPIENT'S RESEARCH PROJECT, as described in Article 3 above.

14. This Agreement will be construed, governed, interpreted and enforced according to the laws of [member state?. Parties will first strive to settle any disputes amicably before taking legal action. All disputes arising out of or in relation to this Agreement that cannot be settled amicably will be brought before the competent court in the [member state], in the district in which the Provider resides.

15. This Agreement will be binding upon and inure to the benefit of the respective successors and assignees of the Parties hereto. However, RECIPIENT may not assign this Agreement in whole or in part without the prior written consent of the PROVIDER.

16. This Agreement may only be altered or amended by an instrument in writing signed by all of the Parties.

17. If any portion of this Agreement is in violation of any applicable regulation, or is unenforceable or void for any reason whatsoever, such portion will be inoperative, and the remainder of this Agreement will be binding upon the Parties.

18. Both Parties acknowledge that the signatories to this Agreement are authorized representatives of each of the Parties and legally authorized to sign this Agreement.

19. If the lawful performance of any part of this Agreement by a Party is rendered impossible by or as a result of any cause beyond such Party's reasonable control, such Party will not be considered in breach hereof as a result of failing so to perform.

IN WITNESS WHEREOF, the Parties have executed this Agreement, in duplicate originals or as a signed PDF, as of the Effective Date.

For the **PROVIDER**

By: ___ _____
Name:
Title
Date: _____

By: _____
Name:
Title
Date: _____

For **RECIPIENT**,

By: _____
Name:
Title
Date: _____

By: _____
Name:
Title
Date: _____

READ AND ACKNOWLEDGED:

_____
PROVIDER'S SCIENTIST

READ AND ACKNOWLEDGED:

_____
RECIPIENT'S SCIENTIST

ANNEX I

Description of the DATA, methods of transfer and storage, allowed processors

Data subjects The personal data transferred concern the following categories of data subjects:	
Purpose of the transfer(s) The transfer is made for the following purpose:	See Annex II
Categories of data The personal data transferred concern the following categories (types) of data:	NB: All health information qualifies as sensitive data as meant in the field below
Sensitive data (if appropriate) e.g.: •racial or ethnic origin, •political opinions, •religious or philosophical beliefs, •trade union membership, •genetic data, biometric data, •health data, •sex life and sexual orientation	
Method of transfer e.g.: Soft- or hardware encrypted USB drive, database entry such as in Castor, etc.	
Method of data storage and security measures (e.g., method of encoding)	
Authorized processors, if applicable, as indicated in clause 3 of the Agreement	

ANNEX II

Recipient's research plan

The research plan sets out the specific purpose(s) for which the DATA may be used.

-

10.3 Letter of Commitment VPHi

Liesbet Geris
VPHi Director
VPH Institute
September 11th 2023

Dear VITO,

Re: Letter of Commitment – REALM Access to Digital Twin (DT)

I am writing on behalf of the VPH institute, to express our commitment to providing access to our wide network of institutions, generating state-of-the-art models and simulation platforms for a range of systems, organs and diseases.

We acknowledge the importance of the collaboration and our promised contribution to the REALM project. We recognize the value of enabling REALM, to critical validated simulation platforms and models to evaluate novel healthcare software across a realistic in-silico population of individual values for target parameters of interest as described in the deliverable D4.5 (Authorized access agreements with European Medical Data Pools and Digital Twin).

In light of this commitment, we hereby confirm the following terms and conditions:

1. Linking to relevant VPHi partners with candidate validated in-silico models or simulation platforms: we commit to contribute to grant REALM access to the DT as described in the deliverable D4.5 (deadline 30th of September) in order to provide on-demand synthetic data (generated with DT) for the evaluation of medical software through REALM.
2. Compliance with Legal and Regulatory Requirements: we will adhere to all applicable legal requirement in facilitating access to the VPHi partners with validated in-silico models of simulation platforms. We will ensure the compliance with data owners (not include in the REALM data pools) used as input by DT. Finally, we will contribute to implement security measures to share synthetic and realistic data with REALM.
3. Support and Training: Provide necessary support and training to setup the suggested technologies for REALM and make it part of the evaluation workflow.

We will proceed with the necessary steps to facilitate access to above mentioned technologies promptly.

Sincerely,

Prof. Liesbet Geris
VPHi Director
VPH Institute
September 11th 2023

Funded by
the European Union

11 Supplement

- Full data poll list (Excel spreadsheet): *Data_pools_list.xlsx*