Project Acronym:	HosmartAI
Grant Agreement number:	101016834 (H2020-DT-2020-1 – Innovation Action)
Project Full Title:	Hospital Smart development based on AI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016834

DELIVERABLE

D6.8 – Data Management Handling Plan – Second version

Dissemination level:	PU -Public
Type of deliverable:	ORDP -Open Research Data Pilot
Contractual date of delivery:	31 October 2022
Deliverable leader:	EXYS
Status - version, date:	Final – v1.0, 2022-10-31
Keywords:	Data management handling plan, Data management strategy, Data management survey, Data storage, Data processing, FAIR, Quality control, Governance, Data security, Research Data Alliance.

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 101016834. The content of this document reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Executive Summary

This deliverable consists of the second version of the DMP (Data Management Handling Plan) for the HosmartAI project, which is funded by the European Union's H2020 Programme under Grant Agreement No. 101016834. The DMP aims to include among other (open) data sources, (open) source code, scientific publications, project deliverables and more. The DMP is used internally by the consortium partners for the effective Data Management Handling (M2-M41, Leader: EXYS). Moreover, the DMP is a mandatory requirement for the HosmartAI project to participate in the Horizon 2020 Open Access to scientific peer-reviewed publications and research data. The second version of the DMP is due by M22.

The HosmartAI project participates in the Pilot on Horizon 2020 Open Research Data. The use of a Data Management Plan is required for all EU H2020 funded projects.

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the Consortium with regard to the project research data.

It also reflects the current state of the Consortium agreements on data management and must be consistent with the exploitation and IPR requirements. Research data linked to exploitable results will not be put into the open domain if they compromise its commercialisation prospects or have inadequate protection, which is an H2020 obligation. The rest of the research data will be deposited in an open access repository.

The DMP covers the full data management life cycle for the data to be collected, processed and generated by the HosmartAI project. Towards handling research data management, within Publications and Research Data in Horizon 2020, detailing

1. how research data will be handled during & after the project;
2. what data will be collected, processed or generated;
3. what methodology & standards will be applied;
4. whether data will be shared /made open access/ how data will be curated and preserved.

The HosmartAI's DMP is based on the **FAIR principles**[1] (Findable, Accessible, Interoperable, Reusable) [REF-05] and on the Guidelines on Implementation of Open Access to Scientific Publications and Research Data in projects supported by the European Research Council (ERC) under Horizon 2020[2].

The first version of the Data Management Plan (DMP, deliverable D6.7), which context is T6.5, was formulated and delivered in M6 of the project implementation.

The DMP is expected to mature in the course of the project, therefore it constitutes a living document that the involved partner can edit. Eventually, the DMP will be finalised and delivered in its definitive form at the end of the project. EXYS is leading the activity while all data provision partners are investing effort in safeguarding their proper data management.

[1] The FAIR Guiding Principles for scientific data management and stewardship, March 2016 https://www.nature.com/articles/sdata201618

[2] https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf

The first version of the DMP included an overview of the datasets to be produced by the project, and the specific conditions that are attached to them. This second version of the HosmartAI DMP delves into the topics related to the research datasets management, continues the work of filling in data in the tables, and, in particular, collects information about the "Allocation of resources" (Section 4), which was not included in the first version of the DMP. A new section (3.4.4) on RDA FAIR Data Maturity Model implementation guidelines was also added.

In general, the DMP first evaluated the legal frameworks and the requirements for all pilots, then examined whether there are currently available data to which open access can be granted, always respecting the security and privacy requirements imposed. With regards to the dissemination of the scientific results, the consortium will establish and promote open access publications, and partners will be encouraged to publish open access articles, so as to enable researchers to build upon previous research results, foster collaborations, avoid duplication of efforts, and accelerate innovation.

Project members will be offered the option of publishing in journals contained/registered in the ROAR[3] and/or other repositories (OpenDOAR[4], OpenAIRE[5] and Zenodo[6]). Authors' copyright agreements will determine whether scientific publications resulting from the project will adopt the gold or the green model.

In addition to the task leader EXYS, contributors to this deliverable are **TGLV, PhE, UKCM, IRCCS, SERMAS, FIBHULP, CHUL, INTRAS, AHEPA, VUB, UM** and **PHILIPS.**

The structure of the DMP follows the H2020 Data Management Plan template.

Data Management Strategy: to map all relevant data collections and to establish the data management needs, in the first version of the DMP every Pilot leader was requested to complete a 'Data management survey', which was reported in the Appendix B for that deliverable.

As is described in Section 1.1, the HosmartAI project will base its outcomes and improvements on eight large-scale pilots, therefore the research data management and the DMP is largely based on these pilots.

[3] ROAR: Research Open Access Repository, http://roar.eprints.org

[4] OpenDOAR: Directory for Open Access Repositories: https://v2.sherpa.ac.uk/opendoar/

[5] https://www.openaire.eu/

[6] https://zenodo.org /

Deliverable leader:	Angelo Consoli (EXYS)
Contributors:	Luca Gilardi, Angelo Consoli (EXYS) Georgios Rampidis (AHEPA, pilot #1) Nicola Bettin (VIMAR, pilot #3) Arktos Meholli (SERMAS, pilot #4, Chapter 4) Marcela Chavez, Patrick Duflot (CHUL, pilot #2) Izidor Mlakar (UM, UKCM, pilot #5) Rosa Almeida (INTRAS, pilot #6) Diana Marqués (INTRAS, Chapter 4) Carlos Luis Parra-Calderón (EFMI, Section 3.4.4) Nivedita Yadav (VUB, pilots #7 and #8, and Chapter 4) Robert Hofsink (PHILIPS, pilot #7) Magda Chatzikou (PhE) Dieter De Court, Wim Vranken (VUB, pilot #8)
Reviewers:	Robert Hofsink (PHILIPS), Elizabeth Zuurmond (ETHZ)
Approved by:	Athanasios Poulakidas, Irene Diamantopoulou (INTRA)

Document History

Version	**Date**	**Contributor(s)**	**Description**
0.1	2021-06-26	Luca Gilardi	2nd iteration – new live document
0.2	2022-04-27	Arktos Meholli, Luca Gilardi	Added Data for Pilot #4
0.3	2022-08-16	Luca Gilardi, Angelo Consoli	Document extended and information added.
0.4	2022-10-04	Izidor Maklas, Arktos Meholli	Various contributions about pilots, and to Section 4
0.5	2022-10-12	Diana Marqués, Manos Georgoudakis, Patrick Duflot	Contributions to Section 4, and data tables.
0.6	2022-10-18	Carlos Luis Parra Calderon, Nivedita Yadav	New section on RDA FAIR Data Maturity Model guidelines, contributions to pilot 7 and 8, and Section 4.
0.7	2022-10-20	Luca Gilardi	Conclusions and finalization before internal review
0.8	2022-10-20	Angelo Consoli	Pre-review and small refinements
0.9	2022-10-26	Elizabeth Zuurmond	Review from ETHZ
0.10	2022-10-28	Robert Hofsink	Review from PHILIPS
1.0	2022-10-31	Athanasios Poulakidas, Irene Diamantopoulou	QA and creation of the final submitted version

Table of Contents

List of Tables

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	**Title**
ASR	Automatic Speech Recognition
Cath lab	Catheterization laboratory
CCDS	Consensus Coding Sequence
CCTA	Coronary Computed Tomography Angiography
COBIT	Control Objectives for Information and related Technology
CPOE	Computerized Physician Order Entry
CRF	Case Report Form
CSV	Comma-Separated Values
CT	Computed Tomography
DICOM	Digital Imaging and COmmunications in Medicine
DMP	Data Management Plan
DPO	Data Protection Officer
DTA	Data Transfer Agreement
EHR	Electronic Health Record
ERC	European Research Council
FAIR	Findable, Accessible, Interoperable, Reusable
FFRCT	Fractional Flow Reserve derived from CT
FHIR	Fast Healthcare Interoperability Resources[7]
GDPR	General Data Protection Regulation[8]
HL7	Health Level Seven
iFR	instantaneous wave-Free ratio Resting index
IMU	Inertial Measurement Unit
IVUS	Intravascular Ultrasound
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LPD	Swiss Data Protection Law
OCT	Optical Coherence Tomography
PAM	Privilege Access Management
PGHD	Patient Generated Health Data
PREM	Patient Reported Experience Measure
PROM	Patient Reported Outcome Measure
RDA	Research Data Alliance
RFR	Coronary Physiology Resting Full-Cycle Ratio
SOP	Standard Operating Procedure
SPSS	Statistical Package for the Social Sciences[9]
SSH	Secure Shell

[7] HL7 FHIR (Fast Healthcare Interoperability Resources), http://hl7.org/fhir/index.html

[8] https://gdpr-info.eu

[9] https://www.ibm.com/analytics/spss-statistics-software

Acronym/ Abbreviation	Title
SSL / TLS	Secure Sockets Layer / Transport Layer Security
SUS	System Usability Scale
TAM	Technology Acceptance Model
TBD	To be defined/done
TTS	Text-To-Speech
UEQ	User Experience Questionnaire
VPN	Virtual Private Network

Term	Definition
Accessible	Data is Accessible in that it can be always obtained by machines and humans upon appropriate authorization and through a well-defined protocol.
Cohort	In statistics, marketing and demography, a cohort is a group of subjects who share a defining characteristic (typically subjects who experienced a common event in a selected time period, such as birth or graduation).
Findable	Any Data Object should be uniquely and persistently identifiable
Interoperable	The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort. Data Objects can be Interoperable only if: (Meta) data is machine-actionable, (Meta) data formats utilize shared vocabularies and/or ontologies, (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible.
Pseudoanonymisation	The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, as long as such additional information is kept separately and subject to technical and organizational measures to ensure non-attribution to an identified or identifiable individual[10]
Re-usable	For Data Objects to be Re-usable they should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources. Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation.
Schemaless database	A type of database where each item is saved in its own document with a partial schema, leaving the raw information untouched.

[10] GDPR Article 4(3b): https://www.privacy-regulation.eu/en/article-4-definitions-GDPR.htm

1 Introduction

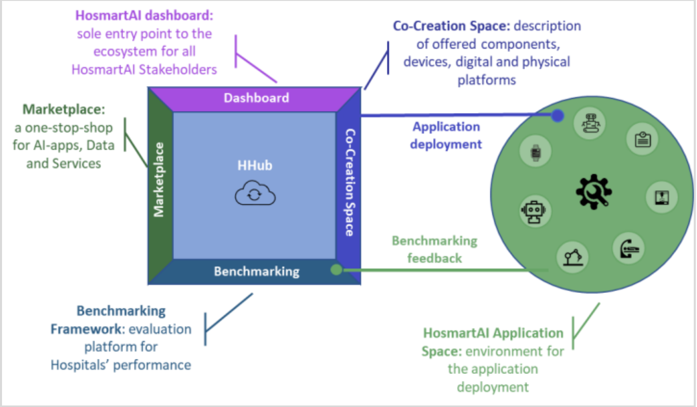
1.1 Project Information

VISION

The HosmartAI vision is a strong, efficient, sustainable and resilient European **Healthcare system** benefiting from the capacities to generate impact of the technology European Stakeholders (SMEs, Research centres, Digital Hubs and Universities).

MISSION

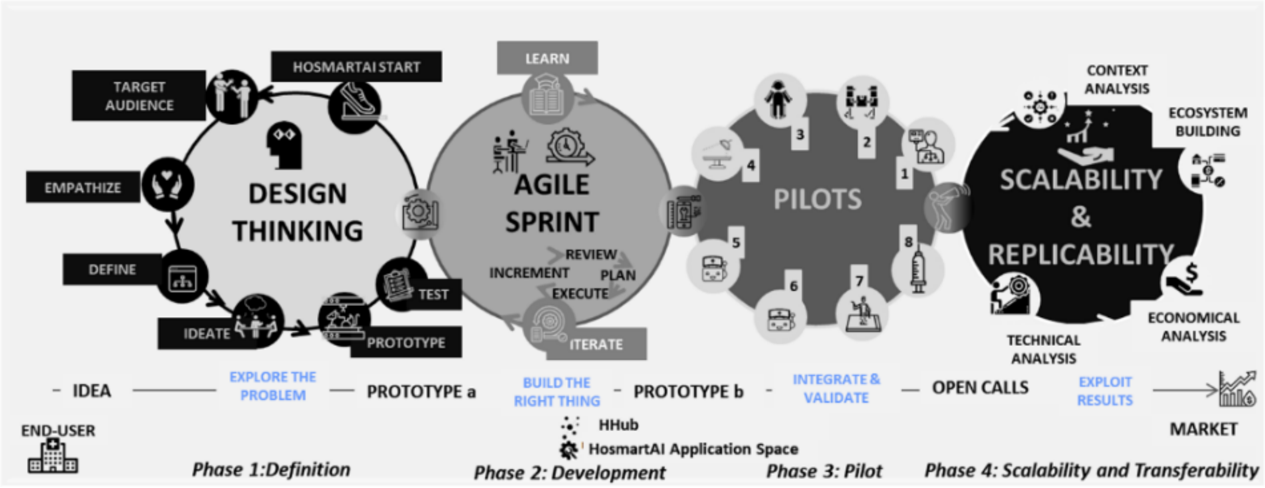
The HosmartAI mission is to guarantee the **integration** of Digital and Robot technologies in new Healthcare environments and the possibility to analyse their benefits by providing an **environment** where digital health care tool providers will be able to design and develop AI solutions as well as a space for the instantiation and deployment of a AI solutions.


HosmartAI dashboard: sole entry point to the ecosystem for all HosmartAI Stakeholders
Co-Creation Space: description of offered components, devices, digital and physical platforms
Marketplace: a one-stop-shop for AI-apps, Data and Services
Dashboard
HHub
Marketplace
Co-Creation Space
Benchmarking
Application deployment
Benchmarking feedback
Benchmarking Framework: evaluation platform for Hospitals' performance
HosmartAI Application Space: environment for the application deployment

HosmartAI will create a common open Integration **Platform** with the necessary tools to facilitate and measure the benefits of integrating digital technologies (robotics and AI) in the healthcare system.

A central **hub** will offer multifaceted lasting functionalities (Marketplace, Co-creation space, Benchmarking) to healthcare stakeholders, combined with a collection of methods, tools and solutions to integrate and deploy AI-enabled solutions. The **Benchmarking** tool will promote the adoption in new settings, while enabling a meeting place for technology providers and end-users.

Eight Large-Scale Pilots will implement and evaluate improvements in medical diagnosis, surgical interventions, prevention and treatment of diseases, and support for rehabilitation and long-term care in several Hospital and care settings. The project will target different **medical** aspects or manifestations such as Cancer (Pilot #1, #2 and #8); Gastrointestinal (GI) disorders (Pilot #1); Cardiovascular diseases (Pilot #1, #4, #5 and #7); Thoracic Disorders (Pilot #5); Neurological diseases (Pilot #3); Elderly Care and Neuropsychological Rehabilitation (Pilot #6); Fetal Growth Restriction (FGR) and Prematurity (Pilot #1).



To ensure a user-centred approach, harmonization in the process (e.g. regarding ethical aspects, standardization, and robustness both from a technical and social and healthcare perspective), the **living lab** methodology will be employed. HosmartAI will identify the appropriate instruments (**KPI**) that measure efficiency without undermining access or quality of care. Liaison and co-operation activities with relevant stakeholders and **open calls** will enable ecosystem building and industrial clustering.

HosmartAI brings together a **consortium** of leading organizations (3 large enterprises, 8 SMEs, 5 hospitals, 4 universities, 2 research centres and 2 associations – see Table 1) along with several more committed organizations (Letters of Support provided).

Table 1: The HosmartAI consortium.

Number[11]	**Name**	**Short name**
1 (CO)	INTRASOFT INTERNATIONAL SA	**INTRA**
1.1 (TP)	INTRASOFT INTERNATIONAL SA	**INTRA-LU**
2	PHILIPS MEDICAL SYSTEMS NEDERLAND BV	**PHILIPS**
3	VIMAR SPA	**VIMAR**
4	GREEN COMMUNICATIONS SAS	**GC**
5	TELEMATIC MEDICAL APPLICATIONS EMPORIA KAI ANAPTIXI PROIONTON TILIATRIKIS MONOPROSOPIKI ETAIRIA PERIORISMENIS EYTHINIS	**TMA**
6	ECLEXYS SAGL	**EXYS**
7	F6S NETWORK IRELAND LIMITED	**F6S**
7.1 (TP)	F6S NETWORK LIMITED	**F6S-UK**
8	PHARMECONS EASY ACCESS LTD	**PhE**
9	SMARTSOL SIA	**TGLV**
10	NINETY ONE GMBH	**91**
11	HEALTH INNOVATION HUB & HOLDING GMBH	**EIT**
12	UNIVERZITETNI KLINICNI CENTER MARIBOR	**UKCM**
13	SAN CAMILLO IRCCS SRL	**IRCCS**

[11] CO: Coordinator. TP: linked third party.

Number[11]	Name	Short name
14	SERVICIO MADRILENO DE SALUD	**SERMAS**
14.1 (TP)	FUNDACION PARA LA INVESTIGACION BIOMEDICA DEL HOSPITAL UNIVERSITARIO LA PAZ	**FIBHULP**
15	CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE	**CHUL**
16	PANEPISTIMIAKO GENIKO NOSOKOMEIO THESSALONIKIS AXEPA	**AHEPA**
17	VRIJE UNIVERSITEIT BRUSSEL	**VUB**
18	ARISTOTELIO PANEPISTIMIO THESSALONIKIS	**AUTH**
19	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	**ETHZ**
20	UNIVERZA V MARIBORU	**UM**
21	INSTITUTO TECNOLÓGICO DE CASTILLA Y LEON	**ITCL**
22	FUNDACION INTRAS	**INTRAS**
23	ASSOCIATION EUROPEAN FEDERATION FORMEDICAL INFORMATICS	**EFMI**
24	FEDERATION EUROPEENNE DES HOPITAUX ET DES SOINS DE SANTE	**HOPE**

1.2 Document Scope

This deliverable aims at deepening the investigation about the research datasets management, continuing the work to collect all data that will be handled by consortium partners in the frame of the HosmartAI project, in particular filling-in the “Allocations of resources” (Section 4), which was missing in the first version of the DMP (D6.7), and providing a new section (3.4.4) about Research Data Alliance FAIR Data Maturity Model implementation guidelines

The initial phase of data collection was followed by the elaboration of the first release of this document, and of its submission on M6. This deadline was a mandatory requirement by the **Horizon 2020 Open Access Data Management** requirement: once a project has had its funding approved and has started, it must submit a first version of the DMP (as a deliverable) within the first 6 months of the project[12]. Other document updates were planned during the project duration, such as the completion of the datasets’ missing information and the allocation of resources for FAIR data management; this allows for accommodating new findings and aligning the DMP to the needs of HosmartAI respecting active regulations.

1.3 Document Structure

This document is comprised of the following chapters:

Chapter 1 is an introduction to the project, the document scope and structure, and gives an overview of the EU-H2020 Open Access programme and of the FAIR principles for data management applied to the HosmartAI project.

Chapter 2 concerns data summary, quotes the Data Management Survey that served as the first version of the DMP, and presents the project’s pilots and purpose of the data collected, the

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

datasets base information, the data types and formats, the physical location of the datasets, their expected sizes, as well as the purposes of the data (data utility) and identification.

Chapter 3 is based on the FAIR principles, and reports about findability of data, provisions for metadata, open accessibility, data interoperability, data reuse through licensing and quality assurance, as well as data cleansing, transforming and analysing.

Chapter 4 illustrates the allocation of resources for data management, i.e. the costs for data FAIR and open access in the HosmartAI project, data management responsibility, costs for long term preservation.

Chapter 5 concerns the provisions for data security and governance, reporting also the COBIT classification degree for every technical partner.

Chapter 6 presents the ethical and legal aspects linked to data management and data sharing.

Chapter 7 illustrates other issues about data management.

Chapter 8 presents tools and references involved in the management of data.

Chapter 9 lists the references of this document.

Chapter 10 presents the conclusions of this document.

Appendix A reports all data tables related to the datasets.

1.4 Open Access

The European Union (EU) strives to improve access to scientific information and to boost the benefits of public investment in research funded under the EU Framework Programme for Research and Innovation Horizon 2020.

Launched by the European Commission along with the H2020 programme, **Open Access** is the practice of providing online access to scientific information that is free of charge to the reader and is reusable. In the context of research and innovation, scientific information can refer to peer-reviewed scientific research articles or research data.

According to this strategy, in Horizon 2020, a limited pilot action on open access to research data has been implemented so that participating projects are required to develop a Data Management Plan (DMP), in which they specify what data will be open.

According to the H2020 **Article 29.2 of the Model Grant Agreement**[13] [REF-01][REF-04], each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results. These open-access requirements are based on a balanced support to both 'Green open access' (immediate or delayed open access that is provided through self-archiving) and 'Gold open access' (immediate open access that is provided by a publisher). Apart from open access to publications, projects must also aim to deposit the research data needed to validate the results presented in the deposited scientific publications, known as "underlying data". In order to

[13] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=242

effectively supply this data, projects need to consider at an early stage how they are going to manage and share the data they create or generate.

Nevertheless, data sharing in the open domain can be restricted if there is a legitimate reason to protect results that can reasonably be expected to be commercially or industrially exploited. Strategies to limit such restrictions will include anonymising or aggregating data, agreeing on a limited embargo period or publishing selected datasets.

From the Data Management point of view, Horizon 2020 strongly suggests that its beneficiaries would make their research data findable, accessible, interoperable and reusable **(FAIR)** to ensure that it is well managed. Good research data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse. The HosmartAI's DMP follows the FAIR principles [REF-02][REF-03], in particular in the health context [REF-07][REF-08].

1.5 FAIR data concepts

The following two subsections report the FAIR main concepts as illustrated in the FORCE11's Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version B1.0 [REF-05].

By adopting all FAIR facets, Data Objects become fully: Findable, Accessible, Interoperable, and Reusable.

The FAIR Data principles aim to ensure that data are shared in a way that enables and enhances reuse by humans and machines. Although FAIR (Findable, Accessible, Interoperable, Reusable) emerged from a workshop for the life-science community, the principles are intended to be applied to data and metadata from all disciplines. Since the formal release via the FORCE11 community, the FAIR data principles have been adopted by several funders and governments worldwide. The European Commission data management guidelines were updated in 2017 to introduce the concept of FAIR [REF-05].

1.5.1 Definitions

- A **Concept** is any defined 'unit of thought' to which we refer in our digital formats.
- A **Data Object** is defined for the purpose of the principles below as: An **Identifiable Data Item with Data elements + Metadata + an Identifier**.
- When we use the term (**Meta) data**, we intend to indicate that the principle is true for **Metadata** as well as for the actual, collected **Data Elements** in the Data Object, and that the principle in question can be independently implemented for both.

1.5.2 FAIR Guiding Principles

1. To be **Findable** any Data Object should be uniquely and persistently identifiable.
   1.1. The same Data Object should be re-findable at any point in time; thus, Data Objects should be **persistent**, with emphasis on their metadata.

   1.2. A Data Object should minimally contain basic machine-actionable metadata that allows it to be distinguished from other Data Objects.
   1.3. Identifiers for any concept used in Data Objects should therefore be Unique and **Persistent**.
2. Data is **Accessible** in that it can always be obtained by machines and humans.
   2.1. Upon appropriate authorization.
   2.2. Through a well-defined protocol.
   2.3. Thus, machines and humans alike will be able to judge the actual accessibility of each Data Object.
3. Data Objects can be **Interoperable** only if:
   3.1. (Meta) data is machine-actionable.
   3.2. (Meta) data formats utilize shared vocabularies and/or ontologies.
   3.3. (Meta) data within the Data Object should thus be both syntactically parseable and semantically machine-accessible.
4. For Data Objects to be **Re-usable** additional criteria are:
   4.1. Data Objects should be compliant with principles 1-3.
   4.2. (Meta) data should be sufficiently well-described and rich that it can be automatically (or with minimal human effort) linked or integrated, like-with-like, with other data sources.
   4.3. Published Data Objects should refer to their sources with rich enough metadata and provenance to enable proper citation.

In order to know the maturity level of the application of FAIR principles in this DMP, each pilot, as well as the entire platform, will be checked using the specifications and guidelines of the Research Data Alliance (RDA) FAIR Data Maturity Model (https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0).
See Section 3.4.4.

2 Data summary

As a first step, the following questions inherited from the indications of H2020 Data Management directives were addressed:

- What is the purpose of the data collection/generation and its relation to the objectives of the project?
- What types and formats of data will the project generate/collect?
- Will you re-use any existing data and how?
- What is the origin of the data?
- What is the expected size of the data?
- To whom might it be useful ('data utility')?

2.1 Data Management Survey

For the purpose of collecting information about the research data processed in the frame of the project, a Data Management Survey was developed and proposed to the pilot leaders and technical partners in the first 6 months from project initiation. The involved partners were:

AHEPA	for pilot #1
CHUL	for pilot #2
IRCCS	for pilot #3
SERMAS and FIBHULP	for pilot #4
UM and UKCM	for pilot #5
INTRAS	for pilot #6
PHILIPS	for pilot #7
VUB	for pilot #8
TGLV	for various research data
PhE	for various research data

The information collected in the survey was used to complete the first version of the DMP.

Important notice:

This document is the second of a series of three deliverable documents (D6.7, D6.8 and D6.9) planned during the project. It is a living document and as such the same will be subject to updates during the HOSMARTAI project. D6.8 is based on the outcomes of D6.7, which in turn was founded on a survey carried out during the first five months of the project. All 8 pilots have contributed by this stage, in particular with information about the allocation of resources and costs in pilot sites. Some aspects related to data management still remain to be specified, and the missing objectives will be fulfilled in the last release of the DMP (D6.9).

2.2 Purpose of the data (HosmartAI pilots)

As explained in the introduction (Section 1.1), eight large-scale pilots are being implemented in the HosmartAI project: improving medical diagnosis, surgical interventions, prevention and treatment of diseases, and support for rehabilitation and long-term care in several hospitals and care settings. These pilots are targeting several **medical** aspects.

Table 2 gives an overview of the eight pilots in relation to the technologies involved.

Table 2: The HosmartAI pilots.

Pilot #	Pilot title	Domain	Main technologies involved	Site	Pilot leader
1	Development of a clinician-friendly, interpretable computer-aided diagnosis system (ICADx) to support and optimise clinical decision making in multi-specialty healthcare environment.	Diagnosis Revolution	Computer-aided diagnosis system.	AHEPA Hospital & Hippokrateio General Hospital of Thessaloniki (Greece)	AHEPA
2	Optimizing the use of radiotherapy	Logistic Improvement	AI algorithm for optimizing patient scheduling.	CHUL Hospital (Belgium)	CHUL
3	Treatment Improvement with the use of innovative technologies and robotics in rehabilitation process	Treatment Improvement	Centralized data collection from wearable devices and environmental sensors.	IRCCS Rehabilitation Centre (Italy)	IRCCS
4	Robotic Systems for minimally Invasive Operation	Surgical support	Robotic system for cardiac catheter navigation, AI and Big Data techniques.	SERMAS Hospital (Spain)	SERMAS
5	Assistive Care in Hospital: Robotic Nurse	Assistive Care	Robotic nurse and Integration of data measured with digital devices.	UKCM Hospital (Slovenia)	UM
6	Assistive Care in Care Centre: Virtual Assistant	Assistive Care	Socially Assistive Robots, eCoach.	INTRAS Care Centre (Spain)	INTRAS
7	Smart Cathlab Assistant	Surgical Support	AI-enabled tools to provide real-time	UZ Brussel (Belgium)	PHILIPS/UZ Brussels

Pilot #	Pilot title	Domain	Main technologies involved	Site	Pilot leader
			clinical decision support and to alleviate the administrative burden in the interventional suite.		
8	Prognosis of cancer patients and their response to treatment combining multi-omics data	Diagnosis and treatment Improvement	General framework to store and analyse raw medical data.	UZ Brussel (Belgium)	VUB

The main purpose of HosmartAI's pilots and the related data outcoming from the HosmartAI project is to improve efficiency in several areas of the medical field (as mentioned before). Moreover, data and datasets are intended to be made open to researchers in the field.

2.3 Datasets, base information

Base information of HosmartAI datasets consists of the type of data to be collected, the name of the datasets, the related pilots and tasks, as well as the responsible and collaborating partners in charge of managing the different datasets handled by the project.

Reference: Appendix A.1

2.4 Data types and formats, physical location

This section reports the actual physical location where the original datasets will be stored. Moreover, it gives the software tools used for data storage and the data standards used for storing the datasets.

Reference: Appendix A.2

2.5 Expected sizes and data volumes (nr. of records)

This section gives the expected size of the datasets used in the HosmartAI project.

Reference: Appendix A.3

2.6 Data utility and identification

Purpose of data and how they can be re-used (data utility) are important aspects of data management for this project. Reported here is the identifiability of collected information for each dataset and how the data will be made accessible to the consortium partners (and whether any restrictions apply).

Reference: Appendix A.4

3 FAIR

3.1 Making data findable, including provisions for metadata

This section shows if data are discoverable with metadata, identifiable and locatable by means of a standard identification mechanism, reports the identified naming convention, and the search keywords (if any) and the datasets version number to optimize the re-use.

Reference: Appendix A.5.

3.2 Making data openly accessible

All aspects of data open accessibility are covered in this section. This includes which datasets are to be made openly available and in which open repositories those will be hosted, as well as the licenses accompanying them, the access identification and the possible restrictions.

Reference: Appendix A.6.

3.3 Making data interoperable

To identify the interoperability of the HosmartAI datasets, metadata vocabularies, ontologies, standards and general methodologies for data interoperability are provided.

Reference: Appendix A.7.

3.4 Increase data re-use (through clarifying licenses)

Re-use of data from the project’s datasets will be accomplished by answering the following questions:

- How will the data be licensed to permit the widest re-use possible?
- When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.
- Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the reuse of some data is restricted, explain why.
- How long is it intended that the data remains re-usable?
- Are data quality assurance processes described?

3.4.1 Data licensing, availability and usability by third parties

This subsection reports the availability, usability and licensing of the project data towards third parties.

Reference: Appendix A.8.

3.4.2 Data Quality Assurance processes

Data Quality Assurance is achieved by specifying for each pilot and task number/dataset the types of analysis that will be performed, the person in charge of creating the statistical analysis plan, and how the transformations and analyses on the data will be verified.

Reference: Appendix A.9.

3.4.3 Data cleansing, transforming and analysing

Several post-processing activities are envisaged on the project's datasets. Information on data cleansing, transforming and analysing, including the type of data cleaning needed (e.g. correct data types, duplicate removal, add missing info…), the person responsible for data cleaning, the type of data transformation/analysis (e.g. normalization, discretization, …), the software/tools used for cleaning, transforming, and analysing, where and by whom the analysis will be conducted, and finally, the standards followed for code development/access and re-use (if available).

Reference: Appendix A.10.

3.4.4 Research Data Alliance (RDA) FAIR Data Maturity Model Implementation Guide

Since Task 6.3 regarding legislation and standardization, work is being done on the application of the RDA's FAIR Data Maturity Model [REF-10].

The RDA FAIR Data Maturity Model Working Group has delivered a set of indicators with priorities and guidelines that provide a 'lingua franca' that can be used to make the results of the assessment using those methodologies and tools comparable. The model can act as a tool that can be used by various stakeholders, including researchers, and data stewards,

The model can act as a tool to be used by various stakeholders to increase the potential for the reuse of research data.

The indicators that are used in the FAIR Data Maturity Model are derived from the FAIR principles and aim to formulate measurable aspects of each principle that can be used by evaluation approaches

As the scientific community values datasets that underpin research findings, the need to ensure the quality, understanding, and consistency of how these datasets are prepared for others to discover and experience requires a method of measurement. The FAIR Data Maturity Model provides a way for these community-based FAIR assessments to have comparable results and provide consistent feedback as to how well communities are doing in making research data FAIR.

Creating FAIR and AI-ready datasets is transforming the state of AI research practice across disciplines. In light of these activities, and given the growth and impact of AI programs for science, it is critical to define at a practical level what FAIR means for AI models. We do this because there is an agreed set of guidelines to FAIRify scientific datasets, from which we will define a compliance model with practical FAIR principles for AI models.

The concept of profiles is being used in the FAIR profile development and adoption community for the practical implementation of FAIR principles in specific domains while ensuring convergence towards a universal model of adoption of the FAIR principles (FAIR convergence).

It is not used to define adaptations of the RDA maturity model, which is absolutely generalist. We propose an implementation guide of the FAIR data maturity model oriented to facilitate adoption and to infer the implementation profile, promoting the development of an implementation profile of FAIR principles in HosmartAI to drive the application of AI.

In the domain of data collected by IoT devices in Healthcare, the aim is to develop a profile that facilitates the sharing and reusability of the data sets, extending discoverability even when combined with the use of AI methods.

A later version of our profiling proposal will consider the FAIR implementation profiling methods being proposed from Go-FAIR, which attempt to maintain the convergence of FAIR methods across different domains [REF-10].

It is intended to implement the two profiles through the development of specific workshops in collaboration with the HosmartAI pilots, within the timeframe of Task 6.3 development (M5-M41).

4 Allocation of resources

This section reports the expected costs for making data FAIR in the HosmartAI project and for long term preservation, and how these costs will be covered (taking into consideration that the costs related to open access to research data are eligible as part of the Horizon 2020 grant). The responsible partner for data management is also indicated, when applicable.

4.1 Costs of FAIR and non-FAIR data in HosmartAI

Table 3: Costs for making data FAIR.

Pilot #	Expected costs for making data FAIR
1	Still TBD[14]
2	MosaiQ Machine. The relevant data are free by DICOM port for ELEKTRA. Electronic Health Record. History of patient (OMNIPRO). TMA: Data is extracted from CHUL data source through REST queries. Data is transformed to FHIR model with combination of open source / proprietary software. Data is stored to FHIR server.
3	Still TBD
4	Still TBD
5	Retrospective data will not be shared. The cost of transforming data to FAIR would include: 1. Extraction of data from the specific clinical cohorts (manual extraction via queries) 2. Transformation to FHIR model (automatic, proprietary software) 3. Iterative process (half-automated): Verification of level of anonymity and anonymization (i.e. removal of risks related to re-identification). Open-source tools can be used to decrease the cost. Prospectively collected data can be shared under FAIR conditions. The cost of transforming data to FAIR will include: 1. Extraction of data from the specific clinical cohorts (automatic, since prospective Patient generated health data – PGHD - will be stored in FHIR) 2. Creation of meta description for the dataset for the given cohort. Open tools can be sued to partially automate the process. 3. Iterative process (half-automated): Verification of level of anonymity (half-automated) and anonymization (i.e., removal of risks related to re-identification). Open-source tools can be used to decrease the cost. 4. Storage of datasets. The use of predefined cohorts published on an open repository (e.g. EOSC) can further mitigate the cost.
6	INTRAS: - Extraction of Gradior data, both patient and outcome data will be done semi-automatically following a convertible scheme to FHIR.

[14] Still To Be Done: not yet established, because some pilots are still in a phase of adaptation and tuning; these data will be decided and eventually reported in definitive form in the final version of the DMP (deliverable D6.9).

	- Transformation to FHIR model (automatic, https://hapifhir.hhub.hosmartai.eu/auth). AUTH: Data collected prospectively using iPrognosis tools could be shared under FAIR conditions. The cost of transforming data to FAIR would include: 1. Extraction of data from the storage databases (hosted either on the cloud or on the pilot's infrastructure). 2. Transformation to the FHIR model. 3. Creation of meta description of the dataset(s). 4. Storage of the dataset(s). TMA: – Extraction data from E-pokratis sensors – Mapping data – Conversion to FHIR model Storage to FHIR server
7	Still TBD
8	Sufficient meta data labelling needs to be done. For genetic and imaging data, it is FAIR as the data is mostly pulled from public databases. We have to make the data findable via search engine. Patient data from UZ Brussel hospital, cancer variant Database that is similar to rare diseases at the moment. We need to incorporate this data in EU health data space[15] for that inter operatable ontology needs to be defined GONE. Reproducible is not possible as it's a patient data. It should not cost anything as it is done at hospital level. We are ready with all the information we can provide to go into the EU health data space.

4.2 Responsible partner for data management

Table 4: Data management responsible.

Pilot #	Data management responsible partner
1	Still TBD
2	TMA: Collect and store data from the machine. Patient is the data owner, ITCL is the data controller while TMA is the data processor. TMA performs data pre-processing and validation. CHUL: Generate anonymized datasheet.
3	Still TBD
4	Still TBD
5	Patient is the data owner, UKCM is the data controller, UM, ITCL and GC are the data processors. UKCM retrieves patient consent, exports the data and defines the conditions regarding the level of anonymization and use. There is a pre-established DTA signed between Izidor Mlakar (UM) and UKCM to allow access to clinical (retrospective) data. UM and UKCM anonymize the data and ensure the level of anonymization is maintained.

[15] https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space_en

6	INTRAS: responsible for the data collected from Gradior sessions on its servers as well as for generating the anonymized data for sending to the HosmartAI platform. AUTH: responsible for handling the data collected from iPrognosis tools (performing any processing required by the iPrognosis tools and/or sending it to storage infrastructure). TMA: responsible for data collection from E-pokratis sensors, validation check and pre-processing (convert to FHIR model and store).
7	Data controllers: Jean-francois Argacha (UZ Brussel), Bert Vandeloo (UZ Brussel) and senior cathlab staff (UZ Brussel). Data processor: Jean-francois Argacha (UZ Brussel), Bert Vandeloo (UZ Brussel), UZ Brussels IT, Philips IT specialist in charge of the study.
8	All relevant personal and clinical data of the participating patients will be processed by the PI of this study or anyone working directly under his supervision. Only those investigators with a medical background (physicians) and physicians caring for the patient, will have access to personal and clinical data. All other involved parties and investigators will only be able to view anonymized data.

4.3 Costs for long term preservations

Table 5: Long term preservation costs.

Pilot #	Long term preservation costs
1	Still TBD
2	CHUL hospital has to analyse if the management of the anonymized datasheet will have long term preservation costs.
3	Still TBD
4	Still TBD
5	The process of exports of anonymized and full de-identifiable data is done per request and is not stored separately by the UKCM. In HosmartAI, the PGHD (pseudo anonymized and fully identifiable) will be stored on a dedicated on-site FHIR server in a private area network (PAN). The access to the server will be IP and MAC restricted. HosmartAI KPIs will be stored in the HosmartAI FHIR server (this is fully anonymized data). The long term-preservation costs include the allocation of a new FHIR server infrastructure (6-10k euro) and operational costs, and costs related to the maintenance of the server (2-4k euro per year).
6	Current legislation requires that personal information included in this study be kept for 5 to 10 years.
7	Data sent outside UZB to Philips will be pseudonymized. Patient administrative data will be deleted.
8	Current legislation requires that personal information included in this study be kept for 20 to 30 years if this data is also part of their medical record at UZB hospital. The data will be stored at the hospital level.

5 Data security

This section aims to answer the following questions:

- What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?
- Is the data safely stored in certified repositories for long term preservation and curation?

5.1 Provisions for data security and governance

Classification

For the assessment of governance and security, the maturity of the process is classified in maturity levels using values that are inherited from the COBIT standardization guidelines[16]. Those values are:

Table 6: Report on the COBIT classification degree for data security provision for every technical partner.

Acronym	Name	Description
I	**Initial**	Process not specified, based on spontaneous initiative that is poorly controlled and reactive.
M	**Managed**	Process is planned, documented and monitored at the project level but not integrated in a broader scope at organization level.
D	**Defined**	Proactive process active at organization level.
Q	**Quantitatively Managed**	The process is measured and controlled/ verified.
O	**Optimizing**	Focus is on continuous process and improvement.

Table 7: Data security.

Partner short name	Classification	Code of conduct used for each task	Levels of data security in place.	Levels of security that primary data used in the project will undergo	Name and email of the data privacy officer(s)
EXYS	Q	Data handling procedures at EXYS are defined according to GDPR and LPD (Swiss data protection law)	Technical measures: • Data storage in dedicated servers segmented on network environment • Authorized access with data access audit logs • VPN regulated access to the data processing data centre • Data communication	Whenever required by the project, data will be stored in firewall-protected computers with strong authentication.	Angelo Consoli; angelo.consoli@eclexys.com

[16] ISACA®, *COBIT® 2019 Framework: Governance and Management Objectives*, USA 2018

Partner short name	Classification	Code of conduct used for each task	Levels of data security in place.	Levels of security that primary data used in the project will undergo	Name and email of the data privacy officer(s)
			and transfer protected by Secure Socket Layer (SSL) and Transport Layer Security (TLS) protocols.		
AUTH	N/A	Data are handled according to GDPR and international ethical guidelines (inter alia, the Word Medical Association Declaration of Helsinki), mandated by the AUTH Research Ethics Committee. Ethical approval must be obtained before data collection involving human beings.	Technical measures include: • Data storage in firewall-protected computers • Authorized access with data access audit logs • Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols.	• Data will be stored in firewall-protected computers with authorized access. • Access will be limited to members of the AUTH research team. • In case of data transfer, this will take place via SSH or SSL protocols. • Pseudo-anonymised data will be stored separately from records including subjects' personally identifiable information (e.g. signed consent forms)	Ms. Kornilia Skarpeta data.protection@auth.gr
VIMAR	M	Data are handled according to GDPR.	Technical measures include: • Data storage in firewall-protected computers. • Authorized access with data access audit logs. • Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols.	• Data will be stored in firewall-protected computers with authorized access. • Access will be limited to authorized members of the team. • In case of data transfer, this will take place via SSL protocols. • Pseudo-anonymised data will be stored separately from records including subjects' personally identifiable information.	Beni Luigi Gianesin beni.gianesin@vimar.com

Partner short name	Classification	Code of conduct used for each task	Levels of data security in place.	Levels of security that primary data used in the project will undergo	Name and email of the data privacy officer(s)
PhE	N/A	Data analysis will be handled according to GDPR following the PhE Standard Operating Procedure of Process & Handling of Personal Data.	Cloud – One Drive	Data will be password protected and accessed only by the PhE HOSMARTAI dedicated team.	George Tsonis Tsonislaw@yahoo.gr
CHUL	Q	Data are handled according to GDPR mandated by both DPO and the CHU Liège Ethics Committee			Ghislaine.Dumont@chuliege.be
ITCL	?	Data are handled according to GDPR and international ethical guidelines. Ethical approval must be obtained before data collection involving human beings.	Technical measures include: • Data storage in firewall-protected computers • Authorized access with data access audit logs • Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols. • Encryption of data.	Data will be stored in firewall-protected computers with authorized access. Access will be limited to members of the ITCL research team. In case of data transfer, this will take place via SSH or SSL protocols. Pseudo-anonymised data will be stored separately from records including subjects' personally identifiable information (e.g. signed consent forms) Critical data will be encrypted for better patient safety.	Angel Lopez, angel.lopez@itcl.es
VUB	I/M	Data are handled according to GDPR and international ethical guidelines, as well as internal UZ Brussel patient data security rules.	The data will remain in the tightly controlled and firewalled UZ Brussel ICT system, with only local or VPN access to limited authorised subsystems.	Limited access by authorized personnel only, no access from outside nor to outside from virtual machine harbouring the data	Data protection office (dpo@vub.be)

Partner short name	Classification	Code of conduct used for each task	Levels of data security in place.	Levels of security that primary data used in the project will undergo	Name and email of the data privacy officer(s)
UZB, VUB	?	Data will be handled according to EU 2016/679 GDPR.	Data storage in firewall-protected computers Authorized access with data access audit logs Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols.	Data will be stored in firewall-protected computers with authorized access. Access will be limited to authorized members of the team. In case of data transfer, this will take place via SSL protocols.	Luc Maes lucm@uzbrussel.be
INTRAS	Q	Data handling procedures at INTRAS are defined according to GDPR	Technical measures: - Data storage in dedicated servers segmented on network environment - Authorized access with data access audit logs - VPN regulated access to the data processing data centre Data communication and transfer protected by Secure Socket Layer (SSL) and Transport Layer Security (TLS) protocols.	- Data will be stored in firewall-protected computers with authorized access. - Access will be limited to members of the INTRAS research team. - In case of data transfer, this will take place via SSH or SSL protocols. Pseudo-anonymised data will be stored separately from records including subjects' personally identifiable information (e.g. signed consent forms)	Francisco Mallo Vázquez <fmv@intras.es>
UM	N/A	Data are handled according to UM's privacy policy (https://feri.um.si/en/about-us/privacy-policy/)	Technical measures include: • Data storage in firewall-protected computers • Authorized access with data access audit logs • Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols	• UM will not store sensitive data	Doc. Dr. Miha Dvojmoč (dpo@um.si), if contacted please always refer project and the main contact person of the project (izidor.mlakar@um.si)

Partner short name	Classification	Code of conduct used for each task	Levels of data security in place.	Levels of security that primary data used in the project will undergo	Name and email of the data privacy officer(s)
UKCM	M	Data are handled according to GDPR and international ethical guidelines (inter alia, the Word Medical Association Declaration of Helsinki), mandated by the UKCM's and National Ethics Committee. Ethical approval must be obtained before data collection involving human beings.	Technical measures include: • Data storage in firewall-protected computers. • IP and MAC restricted access • Authorized access with data access audit logs • Data transfer via Secure Shell (SSH) or Secure Socket Layer (SSL) protocols.	• Data will be stored in firewall-protected computers with authorized access. • Access will be limited to members of the UKCM and UM research team. A specific DTA will be established. • In case of data transfer, this will take place via SSH or SSL protocols. • Pseudo-anonymised data will be stored separately from records including subjects' personally identifiable information (e.g. signed consent forms)	mag. Klara Mihaldinec, (dpo@ukc-mb.si)
91	Q	Data handling procedures are defined according to HIPAA and GDPR	Data storage in HIPAA compliant cloud, high-end encryption with endpoint threat detection and multi factor authentication	• Data will be stored and managed according to the highest levels of protection and encryption as required by HIPAA and GDPR, with firewall, restricted access, and endpoint cybersecurity threat detection	Durim Krasniqi Durim@91.life
SERMAS	O	Best Practices set out by data privacy officers	Data storage in servers that require authorization for access.	• Access by a password. • Anonymized data retrieval and transfer.	FUNDACIÓN PARA LA INVESTIGACIÓN BIOMÉDICA DEL HOSPITAL UNIVERSITARIO LA PAZ (FIBHULP). protecciondedatos@idipaz.es

6 Ethical and legal aspects

Ethical and legal aspects which are part of the HosmartAI project are treated in this section. The following questions will be addressed:

- Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and the ethics chapter in the Description of the Action (DoA).
- Is an informed consent for data sharing and long-term preservation included in questionnaires dealing with personal data?

6.1 Ethical issues for data sharing

Data Protection is a fundamental human right, as enshrined in the EU Charter of Fundamental Rights, aimed at providing any individual[17] with control over the way their personal information is collected and used. Article 8(1) of the Charter of Fundamental Rights of the European Union (the 'Charter') and Article 16(1) of the Treaty on the Functioning of the European Union (TFEU) grant everyone the right to the protection of their personal data. Data protection is a central issue for research ethics.

Whenever personal data is collected, there are both ethical and legal obligations to ensure that participants' information is properly protected. This is fundamental to safeguarding their rights and freedoms, and minimising the ethics risks related to the data processing. In HosmartAI data security is provided on all levels. Only authorized users will have access to digital information. The project will adopt recommendations and standards provided by ENISA (European Union Agency for Cybersecurity) [REF-06]. It is the goal of all project partners to mitigate the risk for all participating patients

Publication of Results HosmartAI complies with the highest ethical standards. Researchers, authors, sponsors, editors and publishers all have ethical obligations with regard to the publication and dissemination of the research results.

6.1.1 Ethical review

This subsection deals with the type of ethical review needed for each pilot

Reference: Appendix A.11

6.2 Legal issues for data sharing

The GDPR provides the basic legal framework for personal data processing and therefore data sharing. Particular attention must be paid to research involving sensitive data such as health data, which according to GDPR must not be processed unless the data subject has given explicit consent.

[17] An individual is an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Art. 2(a) EU General Data Protection Regulation (GDPR).

This imposes obligations on researchers to provide research subjects with detailed information about what will happen to the personal data that they collect.

In HosmartAI, all data processing complies with EU law as well as national data laws. It is ensured that any partners, contractors or service providers that process research data at HosmartAI partners' request and on their behalf comply with the GDPR and the H2020 ethics standards. Special attention is given to a good balance between research objectives and the means used to achieve them.

Types of Personal Data Processed During the lifetime of HosmartAI two categories of personal and sensitive data are/may be generated or collected: (1) "Data related to stakeholders" (individuals working for the consortium partners or in any way professionally involved with the project, etc.): Information on these data subjects, such as contact details (e.g. e-mails and names), their signatures, authorship of deliverables, etc. is collected and processed by all partners of the Project. (2) "Patient's health data in pilots" (sensitive data or "special personal data" pursuant to Article 9 GDPR):

Informed Consent When personal data is used, informed consent is the cornerstone of research ethics. The lawful basis for the processing of personal data related to stakeholders, under the GDPR, is that each data subject working for a project partner has given consent to the processing of their personal data (GDPR Article 6 (1)(a)) and that the processing is necessary for the performance of a contract - namely, the data subjects' employment agreements with each project partner (GDPR Article 6(1)(b)). At the end of the project, files containing personal information of data subjects working for Project partners will be maintained by each project partner. Any partner will have the right to continue to maintain its copy of the contact data of employees working for HosmartAI partners unless employees have requested a deletion of the contact data. Mailing lists of the project will be deleted only after the very final payment and assessment from the European Commission. Data subjects' contact details will be shared only with project members and only for the time needed to execute the Grant Agreement and/or complete the project. Authorship information may be made publicly available with the consent of the data subjects once the application becomes publicly or commercially available.

Whenever personal data is collected from patients, the patients' informed consent must be sought by means of a procedure that meets the minimum standards of the GDPR. This requires consent to be given by a clear affirmative act. For consent to data processing to be 'informed', the data subject must be provided with detailed information about the envisaged data processing in an intelligible and easily accessible form, using clear and plain language. The researchers at the pilots will explain to patients (i) what the research is about; (ii) what their participation in the project will entail and what risks may be involved. The partner will give information as to whether data will be shared with, or transferred to, third parties and for what purposes and for how long the data will be retained before being destroyed. The patients will also be informed about the right to withdraw consent or access their data. They will also be told the procedures to follow should they wish to do so. They will also receive information on their right to lodge a complaint with a supervisory authority. The data subjects must also be made aware if data are to be used for any other purposes, and if it is to be shared with research partners or transferred to organisations outside the EU. Records documenting the informed consent procedure will be kept, including the information sheets and consent forms provided to research participants. The consent process(es) and the information provided to the data subjects will cover all the data processing activities related to their participation in HosmartAI. If in the course of the HosmartAI research project, any significant changes to methodology or processing arrangements that have a bearing on the data subjects' rights or the use of their data should occur, the data subjects will be made aware of the intended changes, and their express consent for further use of the data will be sought.

Privacy by Design To innovate ethically and responsibly, researchers and developers apply the concept of 'privacy by design', which provides a framework for focusing the design of systems, databases and processes with respect to data subjects' fundamental rights. A wider concept of 'data protection by design', now included in the GDPR, requires the implementation of appropriate technical and organisational measures to give effect to the GDPR's core data-protection principles. Data protection by design is one of the best ways to address the ethics concerns that arise within a research project. Minimisation of data is essential in this respect. Data processing must be lawful, fair and transparent. It should involve only data that are necessary and proportionate to achieve the specific task or purpose for which they are collected.

Deletion and Archiving of Data Personal data will only be kept as long as is necessary for the purposes for which they are collected, or in accordance with the established auditing, archiving or retention provisions of HosmartAI. As soon as the research data is no longer needed, or subject to an established retention period, the data will be deleted. Data retained for auditing processes will be stored securely and further processed for those purposes only. Research data held in the cloud or by a third-party service provider, will also be held together with any back-ups.

Reuse of Data A potential later use of the HosmartAI platform may permit medical researchers to use data sets for the purpose of conducting medical research. The procedure for this eventuality has not yet been addressed by the project partners. As a result of this effort, ethical and legal considerations may arise with respect to large scale or big data processing, and will be discussed in the last version of the DMP.

7 Other issues about data management

This section reports on other possible issues related to data management in the HosmartAI project. For instance, the use of other national/funder/sectorial/departmental procedures for data management.

Pilot #	Other issues related to data management
1	Still TBD
2	CHUL hospital has to establish and manage the anonymized datasheet.
3	Still TBD
4	Still TBD
5	UKCM prefers the decentralized approach in which data is stored within the pilot site. If sharing of prospective data is established, UKCM must create and anonymize data. Individual DTAs with clearly identified intent for the use of 'raw' data and how data will be handled must be negotiated. Fully anonymized cohorts can be shared openly for research.
6	Still TBD
7	Still TBD
8	Still TBD

8 Tools and references

- The Research Data Alliance provides a Metadata Standards Directory that can be searched for discipline-specific standards and associated tools.
- Research Data Alliance FAIR Data Maturity Model (https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines-0).
- The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of an adequate license for research data.
- Useful listings of repositories include:
  - Registry of Research Data Repositories
  - Some repositories like Zenodo, an OpenAIRE and CERN collaboration), allow researchers to deposit both publications and data, while providing tools to link them.
- Other useful tools include DMP online and platforms for making individual scientific observations available such as ScienceMatters.
- Mosaiq Machine data for radiotherapy by FHIR.
- Omnipro[18]. IT CHUL hospital transfer the anonymized therapy data by FHIR.
- FHIR4FAIR Implementation Guide (HL7 project underway, ballot scheduled for September 2021) (http://build.fhir.org/ig/HL7/fhir-forfair/)
- ROAR: Research Open Access Repository, http://roar.eprints.org
- OpenDOAR: Directory for Open Access Repositories: https://v2.sherpa.ac.uk/opendoar/
- OpenAIRE: https://www.openaire.eu/
- Zenodo: https://zenodo.org/
- Elekta MosaiQ Radiation Oncology: https://www.elekta.com/software-solutions/care-management/mosaiq-radiation-oncology/
- GRADIOR: Computer-based cognitive rehabilitation program.
- EVA Corpus: A Corpus for Analysing Linguistic and Paralinguistic Features in Multi-Speaker Spontaneous Conversations
- HBASE, MONGO
- K-Anonymity: https://github.com/Nuclearstar/K-Anonymity
- ARX - Open Source Data Anonymization Software: https://github.com/arx-deidentifier/arx
- European Open Science Cloud (EOSC): https://eosc-portal.eu/

[18] https://cabinetprive.xperthis.com/omnipro/

9 References

[REF-01]	AGA - Annotated Model Grant Agreement, H2020 Programme, v. 5.2., June 2019
[REF-02]	Template for the Data Management Plan, H2020 Programme
[REF-03]	FAIR Data Management template Summary table, H2020 programme
[REF-04]	H2020 Model Grant Agreement - Article 29.2
[REF-05]	FORCE11's Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing (https://www.force11.org/fairprinciples)
[REF-06]	ENISA: European Union Agency for Cybersecurity (https://www.enisa.europa.eu/)
[REF-07]	Jaime Delgado, FAIR4Health - Report on Security and Privacy in FAIR processes, Horizon 2020 project, grant agreement No 824666
[REF-08]	Jaime Delgado and al., Approaches to the integration of TRUST and FAIR principles, Universitat Politècnica de Catalunya (UPC BarcelonaTECH)
[REF-09]	Research Data Alliance (RDA) FAIR Data Maturity Model, https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[REF-10]	Data Intelligence 2(2020), 158-170. doi: 10.1162/dint_a_00038

10 Conclusions

The second version of the HosmartAI DMP deepens the investigations related to the research datasets management, initiated in the first version of the DMP (D6.7). This document continues the work of filling in data in the tables and sections, and, in particular, collects information about the "Allocation of resources" (Section 4), which was not included in the previous version of the DMP. A new section (3.4.4) on the RDA FAIR Data Maturity Model implementation guidelines was also added: this work will be carried on in Task 6.3, through the implementation of two profiles, and will provide a quantitative self-assessment model for measuring the maturity level of the datasets.

Since the implementation and deployment of some pilots are still in a stage of adaptation and tuning, some information about research datasets is still not finalized, and will be provided in the final version of the DMP (deliverable document D6.9).

Evolution towards the final version: the DMP is a living document and further considerations will be made in subsequent iterations, especially with respect to donated health records and potential for the application to be used as a research platform. Moreover, pilot datasets maturity level will be measured quantitatively using the RDA FAIR Data Maturity Model guidelines, implemented in Task T6.3, eventually leading to the final version of the DMP. Some aspects related to data management are still open for some pilots and will be finalized in the next phases of the project; for this reason, this second version contains a certain number of "Still TBD" that will be clarified during the progress of the activities and will be completed in the last version of the DMP.

Appendix A Datasets

A.1 Datasets base information

Table 8: Datasets base information.

Code	Type of data to be collected / name of the dataset	Pilot Nr.	Task Nr.	Responsible partner	Collaborating partners
DS1.1	Cardiac ultrasound video recordings	1	3.1, 5.2	AHEPA	AUTH
DS1.2	Capsule endoscopy video recordings	1	3.1, 5.2	AHEPA	AUTH
DS1.3	Cardiotocography variables and results, biometric data, medical history data	1	3.1, 5.2	AUTH	N/A
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	1	3.1, 5.2	AHEPA	AUTH
DS2.1	Data related to hours spent by specialists.	2	3.2	CHUL	ITCL/TM
DS2.2	Retrospective patient schedule data and precondition of the treatment	2	3.2	CHUL	ITCL/TMA
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	2	1.2	CHUL	N/A
DS2.4	PROMs/PREMs	2	5.2	CHUL	UM
DS2.5	Prospective patient clinical data	2	5.2	CHUL	ICTL/TMA
DS2.6	Patient personal data (address, preferences, …)	2	5.2	CHUL	ICTL/TMA
DS2.7	Retrospective EHR	2	3.2	CHUL	ICTL/TMA
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	2	3.2	ITCL	ICTL/TMA
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human	3	3.3	IRCCS	VIMAR

Code	Type of data to be collected / name of the dataset	Pilot Nr.	Task Nr.	Responsible partner	Collaborating partners
	presence/access control, Devices' activation and connected loads				
DS3.2	IMU data captured by iPrognosis smartphone application	3, 6	5.2	AUTH	INTRAS
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	3, 6	5.2	AUTH	INTRAS
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	3, 6	5.2	AUTH	INTRAS
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	3, 6	5.2	AUTH	INTRAS
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	6	3.5	INTRAS	ITCL/AUTH
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	4	5.2	SERMAS	91
DS4.2	3D navigation system data (intracardiac signals, geometry)	4	5.2	SERMAS	91
DS5.1	Patient recordings and features extracted from interaction with patients	5	5.2	UM	UKCM, ITCL
DS5.2	Biometric data (e.g., blood pressure and heart rate)	5	5.2	UKCM	UM
DS5.3	Retrospective electronic health records	5	5.2	UKCM	UM
DS5.4	PREMs related to clinical staff (depends on T1.4)	5	5.2	UKCM	UM
DS5.5	PREMs related patients (PAM, SUS-SI/TAM, UEQ)	5	5.2	UM	UKCM
DS5.6	Datasets for facial expression and emotion recognition	5	3.5	UM	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	5	3.5	UM	N/A

Code	Type of data to be collected / name of the dataset	Pilot Nr.	Task Nr.	Responsible partner	Collaborating partners
DS5.8	Datasets for ASR and TTS in French	5	3.5	UM	N/A
DS5.9	EVA Corpus[19], data set of conversational expression	5	3.5	UM	N/A
DS5.10	Video recordings of third persons	5	3.5	ITCL	N/A
DS5.11	Patient's behavioural information	5	3.5	ITCL	UM
DS5.12	Patient's facial information	5	3.5	ITCL, UM	N/A
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	6	3.5	ITCL	N/A
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS7.5	Intravascular imaging data of patient evaluated by either	7	3.6	Cardiology department	Philips Image Guided

[19] https://www.iaras.org/iaras/home/cijc/a-corpus-for-analyzing-linguistic-and-paralinguistic-features-in-multi-speaker-spontaneous-conversations-eva-corpus

Code	Type of data to be collected / name of the dataset	Pilot Nr.	Task Nr.	Responsible partner	Collaborating partners
	OCT or IVUS technique before and after a coronary intervention			of University Hospital Brussels, VUB	Therapy Systems (Philips)
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	7	3.6	Cardiology department of University Hospital Brussels, VUB	Philips Image Guided Therapy Systems (Philips)
DS8.1	Image, gene, phenotype and pathology data for glioma patients	8	N/A	VUB	N/A
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	PhE	5.3	All pilots	PhE

A.2 Data types and formats, physical location

Table 9: Data types, formats and physical location.

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
DS1.1	Cardiac ultrasound video recordings	At the edge, dedicated infrastructure for the pilot	Still TBD	DICOM, HL7-compatible resource
DS1.2	Capsule endoscopy video recordings	At the edge, dedicated infrastructure for the pilot	Still TBD	MP4, HL7-compatible resource
DS1.3	Cardiotocography variables and results, biometric data, medical history data	At the edge, dedicated infrastructure for the pilot	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	At the edge, dedicated infrastructure for the pilot	Still TBD	DICOM partially (others TB
DS2.1	Data related to hours spent by specialists.	Private cloud of the University of Liège. This infrastructure is based on ESX virtualized environment. This infrastructure is certified ISO27001 and ISO9001 annually since 2016.	Still TBD	HL7-compatible resource
DS2.2	Retrospective patient schedule data and precondition of the treatment	As D2.1	Mosaiq OIS	Json
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	As D2.1	Mosaiq OIS	Json
DS2.4	PROMs/PREMs	As D2.1	Questionnaires stored in Excel file	Excel format
DS2.5	Prospective patient clinical data	As D2.1	EHR	Json
DS2.6	Retrospective EHR	As D2.1	EHR	Json

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
DS2.7	Data from radiotherapy services, infrastructure. Patient's satisfaction, preferences....	At the edge, dedicated infrastructure for the pilot	Data Store (HBASE, MONGO, RDBMS)	Media Resource and formats of HBASE, MONGO, RDBMS observations
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices activation and connected loads	On VIMAR cloud infrastructure	InfluxDB	JSON
DS3.2	IMU data captured by iPrognosis smartphone application	Cloud infrastructure for the pilots	Schemaless database	JSON, HL7-compatible resource
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Cloud infrastructure for the pilots	Schemaless database	JSON, HL7-compatible resource
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Cloud infrastructure for the pilots	Schemaless database	JSON, HL7-compatible resource
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Cloud infrastructure for the pilots	Schemaless database	JSON, HL7-compatible resource
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Cloud infrastructure for the pilots	Relational databases	JSON

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	SERMAS	Server (provided by 91)	Still TBD
DS4.2	3D navigation system data (intracardiac signals, geometry)	SERMAS Cloud, Navigation system internal memory	Server (provided by 91)	*.xls or *.xlsx files
DS5.1	Patient recordings and features extracted from interaction with patients	At the edge, dedicated infrastructure for the pilot.	Open FHIR Server	FHIR Media Resource FHIR Observation and FHIR Compositions
DS5.2	Biometric data (e.g., blood pressure and heart rate)	At the edge, dedicated infrastructure for the pilot	Open FHIR Server	FHIR Diagnostic Report, and Observation resource
DS5.3	Retrospective electronic health records	At the edge, interface with existing IT system of the hospital	Proprietary Hl7 Compliant Medis Server	N/A
DS5.4	PREMs related to clinical staff (depends on T1.4)	A GDPR compliant online survey infrastructure (https://1ka.arnes.si/index.php?lang_id=2) or UM's CHATBOT for collecting PROMs/PREMs	SPSS or similar, Open FHIR Server	FHIR Questionnaire, FHIR QuestionnaireResponse, FHIR Composition
DS5.5	PREMs related patients (PAM, SUS-SI/TAM, UEQ)	A GDPR compliant online survey infrastructure (https://1ka.arnes.si/index.php?lang_id=2), Open FHIR Server	SPSS or similar, Open FHIR Server	FHIR Questionnaire, FHIR QuestionnaireResponse, FHIR Composition

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
DS5.6	Datasets for facial expression and emotion recognition	UM's internal infrastructure	N/A	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	UM's internal infrastructure	N/A	N/A
DS5.8	Datasets for ASR and TTS in French	UM's internal infrastructure	N/A	N/A
DS5.9	EVA Corpus, data set of conversational expression	UM's internal infrastructure And CLARIN.SI Repository	N/A	N/A
DS5.10	Video recordings of third persons	Inside the robot hardware or attached devices	N/A	Bytestreams
DS5.11	Patient's behavioural information	At the edge, dedicated infrastructure for the pilot	Data Store (HBASE, MONGO, RDBMS)	To be defined
DS5.12	Patient's facial information	At the edge, dedicated infrastructure for the pilot	Data Store (HBASE, MONGO, RDBMS)	FAPS, AUs
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data	At the edge, dedicated infrastructure for the pilot	Data Store (HBASE, MONGO, RDBMS)	Media Resource and formats of HBASE, MONGO, RDBMS observations
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	UZB, VUB	Electronic health records system branded PRIMUZ	xlsx files
DS7.2	Clinical data of patients scheduled for a coronary	UZB, VUB	Electronic health records	xlsx files

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
	angiogram/coronary intervention @ UZB		system branded PRIMUZ	
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	UZB, VUB	Philips cathlabs over the Philips IntelliSpace CardioVascual (ISCV) Portal	DICOM
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	UZB, VUB	Coroventis platform (RFR/FFR) and Volcano platform (iFR/FFR)	Data extracted in .dat or .xls files
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	UZB, VUB	Abbott OPTIS platform (OCT) (Integrated platform) Volcano platform (IVUS) (non-integrated platform)	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	UZB, VUB	Philips (CT) Heartflow (FFRCT)	Still TBD
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary	Philips AI smart-cathlab prototype working at UZB, VUB	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Physical location where primary (original) data is stored	Software/ tools used for storage of data?	Format and type of data standards used to store the data
	physiology, intravascular imaging and coronary CT data.			
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Separate in-hospital databases, with final central integration on site	PRIMUZ, XNAT	RDMS (various), JSON
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	The original data at each pilot's repository. The analysis of PhE will be stored in cloud space of PhE (One Drive)	R, Stata	Econometric data, *dta, *xlsx,

A.3 Expected data sizes and volumes

Table 10: Datasets expected size.

Code	Type of data to be collected / name of the dataset	Record size and expected data volume (n. of records)
DS1.1	Cardiac ultrasound video recordings	Still TBD
DS1.2	Capsule endoscopy video recordings	Recordings from 60 patients
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Still TBD
DS2.1	Data related to hours spent by specialists.	Still TBD
DS2.2	Retrospective patient schedule data and precondition of the treatment	3 Patients anonymized
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	Data from 5 machines
DS2.4	PROMs/PREMs	Still TBD
DS2.5	Prospective patient clinical data	Still TBD

Code	Type of data to be collected / name of the dataset	Record size and expected data volume (n. of records)
DS2.6	Patient personal data (address, preferences, …)	Still TBD
DS2.7	Retrospective EHR	3 Patients anonymized
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	Still TBD
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Still TBD
DS3.2	IMU data captured by iPrognosis smartphone application	Still TBD
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Still TBD
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Still TBD
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Still TBD
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Up to 10 MB per patient
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	50
DS4.2	3D navigation system data (intracardiac signals, geometry)	50
DS5.1	Patient recordings and features extracted from interaction with patients	1 recording per interaction, assuming 5 minutes of interaction of up to 300 MB per recording. Feature files per recording are 15kb per 5 minutes for the handcrafted features and roughly 2.5GB MB per 5-10 minutes for fully low-level feature extraction
DS5.2	Biometric data (e.g., blood pressure and heart rate)	> 10 kb per FHIR resource
DS5.3	Retrospective electronic health records	Up to 100MB per patient
DS5.4	PREMs related to clinical staff (depends on T1.4)	Up to 2 MB per patient

Code	Type of data to be collected / name of the dataset	Record size and expected data volume (n. of records)
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	Up to 2 MB per patient
DS5.6	Datasets for facial expression and emotion recognition	> 10GB
DS5.7	Datasets for ASR and TTS in Slovenian	Still TBD we estimate at least 20h of speech
DS5.8	Datasets for ASR and TTS in French	Still TBD we estimate at least 20h of speech
DS5.9	EVA Corpus, data set of conversational expression	2GB per annotated 1h recording
DS5.10	Video recordings of third persons	Still TBD
DS5.11	Patient’s behavioural information	Still TBD
DS5.12	Patient’s facial information	Still TBD
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	At least 1 recording per day per patient and per signal
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Around 1100 coronary angiograms including 350 CA with PCI performed in cathlab 3 and 4 between 01/11/2020 and 01/05/2021
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Around 1100 coronary angiograms including 350 CA with PCI performed in cathlab 3 and 4 between 01/11/2020 and 01/05/2021
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Around 1100 coronary angiograms including 350 CA with PCI performed in cathlab 3 and 4 between 01/11/2020 and 01/05/2021
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Around 150 single point and 20 motorized pullbacks of physiological evaluations by FFR/iFR/RFR performed in cathlab 3 and 4 between 01/11/2020 and 01/05/2021
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Around 20 OCT and 10 IVUS performed in cathlab 3 and 4

Code	Type of data to be collected / name of the dataset	Record size and expected data volume (n. of records)
		between 01/11/2020 and 01/05/2021
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	Around 300 CTA performed in patient referred for an invasive angiogram performed in cathlab 3 and 4 between 01/11/2020 and 01/05/2021
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Validation study of the AI prototype effects on key performance indicators (hospital and health care productivity). Prospective cohort of 100 patients.
DS8.1	Image, gene, phenotype and pathology data for glioma patients	50 patients/year
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	Based on input provided by pilots

A.4 Data utility and identification

Table 11: Purpose and re-use of data (data utility) and identification.

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
DS1.1	Cardiac ultrasound video recordings	Development and evaluation of AI-assisted cardiology diagnosis tool	Partially identifiable since it will be interlinked with a specific patient. We could interlink extracted features and medical images/videos.	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Request must be issued by each individual partner to AHEPA.
DS1.2	Capsule endoscopy	Development and evaluation of AI-assisted	Partially identifiable since it will be interlinked with a	Access to pseudo-anonymised data is allowed with no

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
	video recordings	gastroenterology diagnosis tool	specific patient. We could interlink extracted features and medical images/videos.	restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Request must be issued by each individual partner to AHEPA.
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Development and evaluation of AI-assisted diagnosis tool	Still TBD, possibly anonymous data, but interconnected with hospital data in case of needed patient identification	It is not available/accessible outside pilot #1 AUTH.
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Development and evaluation of AI-assisted diagnosis tool	Still TBD, possibly anonymous data, but interconnected with hospital data in case of needed patient identification	It is not available/accessible outside pilot #1 AHEPA/AUTH.
DS2.1	Data related to hours spent by specialists.	Create patient’s schedule and assess patient’s satisfaction	Anonymized	None
DS2.2	Retrospective patient schedule data and precondition of the treatment	Develop AI model and digital twin.	Anonymized	Under the conditions of the CA as assessed in D8.3 SELP Impact Assessment.
DS2.3	Data linked to radiotherapy machines (tumours treatment indication,	Develop AI model and digital twin Validation.	Fully identifiable	None

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
	maintenance, building location)			
DS2.4	PROMs/PREMs	Validation (chatbot)	Anonymized	PROM and PREM questionnaire no restriction, however Still TBD how chatbot data will be shared with partners
DS2.5	Prospective patient clinical data	Validation	Fully identifiable, should not be shared.	Private patient information is not available/accessible outside pilot #2.
DS2.6	Patient personal data (address, preferences, …)	Validation	Fully identifiable, should not be shared.	Private patient information is not available/accessible outside pilot #2.
DS2.7	Retrospective EHR	Develop AI model and digital twin	Anonymized	Under the conditions of the CA as assessed in D8.3 SELP Impact Assessment.
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	Development algorithm for Optimization Scheduler.	Fully identifiable, should not be shared.	Private patient information is not available/accessible outside pilot #2. What we can share in extracted is features + pattern identification outcome since this is anonymized and will be also exploited in publications

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Monitor of the environment	Fully identifiable since it will be interlinked with a specific location	Data belongs to the owner of the location where the home automation devices are installed. The owner defines access to the data
DS3.2	IMU data captured by iPrognosis smartphone application	Remote monitoring of people with Parkinson's disease (PwP) via the iPrognosis application	Pseudo-anonymised	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Data access roles will be granted.
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Remote monitoring of PwP via the iPrognosis application	Pseudo-anonymised	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Data access roles will be granted.
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone	Remote monitoring of PwP via the iPrognosis application	Pseudo-anonymised	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
	virtual keyboard			or patent). Data access roles will be granted.
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Remote assessment of the motor status of PwP via the iPrognosis iMAT application	Pseudo-anonymised	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Data access roles will be granted.
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Develop AI intervention model and Validation	Pseudo-anonymised	Access to pseudo-anonymised data is allowed with no restrictions, unless otherwise decided (e.g., in case of pending publication or patent). Data access roles will be granted.
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	Population characterization	Pseudoanonymized	Data will not be available for no other person outside pilot #4.
DS4.2	3D navigation system data (intracardiac signals, geometry)	Data source for AI analysis	Pseudoanonymized	Data will not be available for no other person outside pilot #4.
DS5.1	Patient recordings and features extracted from	Support for Spoken Language Interaction, Classification of	Fully identifiable, should not be shared. We can	It is not available/accessible outside pilot #5 UM/UKCM. Special

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
	interaction with patients	psychological distress (symptoms of depression) and action recognition	think of sharing extracted features	DTA is signed with UKCM as data owner and UM as data processor prior to the trial. ***What we can share is extracted features + classification outcome since this is anonymized and will be also exploited in publications***
DS5.2	Biometric data (e.g., blood pressure and heart rate)	Impact of PGHD on CCDS Efficiency and improved management of clinical parameters	Partially identifiable since it will be interlinked with a specific patient. We could interlink extracted features and biometric data as a data set.	It is not available/accessible outside pilot #5 UM/UKCM. Special DTA with UKCM as data owner and UM as data processor will be prepared and signed prior to the trial. **We will consider, however, creating datasets, such as, correlating interaction features and monitored biomarkers with moods/emotions/psychological distress, to be shared openly for research**
DS5.3	Retrospective electronic health records	Impact of CPOE on clinical routine, CCDS Efficiency and improved	Fully identifiable since it will be interlinked with a specific patient. Cannot be shared.	It is not available/accessible outside pilot #5 UM/UKCM. DTA with UKCM as data

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
		management of clinical parameters		owner and UM as data processor will be prepared and signed prior to the trial.
DS5.4	PREMs related to clinical staff (depends on T1.4)	Impact of Social robotics system on various aspects of clinical workflow, including staff satisfaction and workload	Anonymized, however, a code will be provided for comparison prior to, during and after the intervention	Statistical data and cohorts are available to be shared with the consortium and wider, for research
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	Impact of Social robotics system on quality of care	Anonymized, however, a code will be provided for comparison prior to, during and after the intervention	Statistical data and cohorts available to be shared with the consortium, and wider, for research
DS5.6	Datasets for facial expression and emotion recognition	Development of sensing AI to collect and classify symptoms of depression from facial expressions, speech and text.	Public dataset, please consider the individual licenses and restrictions	The data is openly available, request must be issued by each individual partner to the specific owner of the data set
DS5.7	Datasets for ASR and TTS in Slovenian	Support for Spoken Language Interaction	Proprietary dataset owned by UM. It is Background we bring into the project.	The data is not publicly open. Access can be granted on an individual basis (bilateral agreements which may include charges)
DS5.8	Datasets for ASR and TTS in French	Support for Spoken Language Interaction	Proprietary dataset owned by UM and public datasets.	Access to public datasets must be managed by

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
			It is background we bring into the project.	individual partners, access to UM’s closed datasets can be granted on an individual basis (bilateral agreements which may include charges)
DS5.9	EVA Corpus, data set of conversational expression	Support for Spoken Language Interaction	Proprietary dataset owned by UM. It is background we bring into the project. It is already publicly available.	Access via CLARIN.SI repository
DS5.10	Video recordings of third persons	Still TBD	Still TBD	Still TBD
DS5.11	Patient’s behavioural information	Still TBD	Still TBD	Still TBD
DS5.12	Patient’s facial information	Still TBD	Still TBD	Still TBD
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	Support for identification of abnormal pattern recognition.	Fully identifiable, should not be shared.	Private patient information is not available/accessible outside pilot #6. What we can share in extracted is features + pattern identification outcome since this is anonymized and will be also exploited in publications

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Retrospective dataset starting from 11/2020 used to develop an AI prototype to alleviate the administrative burden in the interventional suite by an automatic procedure tracking.	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Retrospective dataset starting from 11/2020 used to develop an AI prototype to alleviate the administrative burden in the interventional suite by an automatic procedure tracking.	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Retrospective dataset starting from 11/2020 used to develop an AI prototype to ensure an automatic logging a smart reporting of both imaging and patient X-ray dosimetry and a facilitated coronary angiogram	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
		interpretation by calculation of severity scores		
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Retrospective dataset starting from 11/2020 used to develop an AI prototype to ensure a clinical decision support.	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Retrospective dataset starting from 11/2020 used to develop an AI prototype to ensure a clinical decision support.	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a	Retrospective dataset starting from 11/2020 used to develop an AI prototype to ensure a clinical decision support.	Pseudo anonymization. Data will be anonymized before sending to third party (Philips) but the encrypting key will be stored by	Data will be not available/accessible outside pilot #7

Code	Type of data to be collected / name of the dataset	Purpose and re-use of the data (data utility)	Collected information will be identifiable?	How will the data be made accessible to other partners in consortium? Are there any restrictions?
	coronary intervention		the owner of the data (UZB, VUB)	
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Prospective validation of AI prototype to ensure automatic procedure tracking, an automatic logging, a smart reporting and a help to clinical decision support.	Non anonymized data will be treated on site (UZB, VUB) by AI smart cathlab Philips prototype Data required to be sent for an extern analysis by a third party will be anonymized but the encrypting key will be stored by the owner of the data (UZB, VUB)	Data will be not available/accessible outside pilot #7
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Dataset for highlighting relationships between different data types and identifying tumour type	The data will be identifiable and will not be shared as is, but procedures to make them (partially) available will be pursued	The data is restricted, we will pursue ways to make the data (partially) available to the consortium via pseudo anonymization; work is ongoing with UZ Brussel.
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	To identify the efficiency in terms of cost-effectiveness of the new AI technologies vs. the previous state of the art of all pilot data	Patient level data will be anonymized, only the intervention will be known to the statistician/health economist.	The primary data is the property of each pilot, and the results of the economic/PROM/PREM analysis will be available to all partners. Coding of the economic part is the property of PhE.

A.5 Data findability, metadata

Table 12: Making data findable, metadata.

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
DS1.1	Cardiac ultrasound video recordings	Still TBD	Still TBD	Still TBD	Still TBD
DS1.2	Capsule endoscopy video recordings	Still TBD	Still TBD	Still TBD	Still TBD
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD	Still TBD	Still TBD	Still TBD
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Still TBD	Still TBD	Still TBD	Still TBD
DS2.1	Data related to hours spent by specialists.	By date, by system (legacy vs AI). TBC whether this can be published or not.	N/A	Radiotherapy, throughput, scheduling, satisfaction	Still TBD
DS2.2	Retrospective patient schedule data and precondition of the treatment	Sample dataset.	N/A	N/A	N/A
DS2.3	Data linked to radiotherapy machines (tumours treatment indication,	By site	N/A	N/A	N/A

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
	maintenance, building location)				
DS2.4	PROMs/PREMs	Those data should not be persisted/shared beyond the project's scope.	N/A	N/A	N/A
DS2.5	Prospective patient clinical data	Internal use only	N/A	N/A	N/A
DS2.6	Patient personal data (address, preferences, …)	Internal use only	N/A	N/A	N/A
DS2.7	Retrospective EHR	Sample dataset.	N/A	N/A	N/A
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	Still TBD	Still TBD	Still TBD	Still TBD
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Still TBD	Still TBD	Still TBD	Still TBD
DS3.2	IMU data captured by iPrognosis	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
	smartphone application				
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Still TBD	Still TBD	Still TBD	Still TBD
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Still TBD	Still TBD	Still TBD	Still TBD
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Still TBD	Still TBD	Still TBD	Still TBD
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Still TBD	Still TBD	Still TBD	Still TBD
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	Still TBD	Still TBD	Still TBD	Still TBD
DS4.2	3D navigation system data (intracardiac signals, geometry)	Still TBD	Still TBD	Still TBD	Still TBD
DS5.1	Patient recordings and features extracted from interaction with patients	Patient, and Department, ID (internal hospital ID), Date-Time,	Action Units, Acoustic units and concepts, Linguistic units and concerts, XPOS for	Depressive or Anxiety disorder	Still TBD, depending on the open data repository

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
		Unit ID, Room ID	morphology, ICD-10 for disease classification		
DS5.2	Biometric data (e.g. blood pressure and heart rate)	Patient, and Department, ID (internal hospital ID), Date-Time, Monitor ID	LOINC, ICD-10	LOINC Features, ICD-10 disease, procedure	Still TBD, depending on the open data repository
DS5.3	Retrospective electronic health records	N/A	N/A	N/A	N/A
DS5.4	PREMs related to clinical staff (depends on T1.4)	Staff-ID, Department ID, Date-Time	LOINC, ICD-10 or custom if the tool is not supported	Still TBD	Still TBD
DS5.5	PREMs related patients (e.g. PAM, SUS-SI/TAM, UEQ) and PROs	Patient, and Department, ID (internal hospital ID), Date-Time	LOINC, ICD-10 or custom if the tool is not supported	Still TBD	Still TBD
DS5.6	Datasets for facial expression and emotion recognition	N/A	N/A	N/A	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	N/A	N/A	N/A	N/A
DS5.8	Datasets for ASR and TTS in French	N/A	N/A	N/A	N/A
DS5.9	EVA Corpus, data set of conversational expression	N/A	N/A	N/A	N/A

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
DS5.10	Video recording of third persons	By robot ID and date	Still TBD	Still TBD	Still TBD
DS5.11	Patient’s behavioural information	By patient name or ID, robot ID and date	Still TBD	Still TBD	Still TBD
DS5.12	Patient’s facial information	By patient name or ID	Still TBD	Still TBD	Still TBD
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	Age, Body Height, General status. Heartbeat. Oxygen, Blood pressure, Body temperature, Urine, Glucose, Body weight, Stool.	Heartbeat: X' Oxygen: % Blood pressure: mmHg Body temperature: ºC Urine: chromatic scale Glucose: mg-DL Body weight: Kg Stool: Number of times per day.	Still TBD	Still TBD
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD	Still TBD
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD	Still TBD
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Still TBD	Still TBD	Still TBD	Still TBD
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Still TBD	Still TBD	Still TBD	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Data identification and versioning	Naming conventions	Search keywords for re-use	Metadata
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Still TBD	Still TBD	Still TBD	Still TBD
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Unique IDs via patient numbers, versioning for local systems exists. Still TBD for the central one.	Still TBD	Still TBD	Still TBD

A.6 Data open accessibility

Table 13: Making data open accessible.

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
DS1.1	Cardiac ultrasound video recordings	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS1.2	Capsule endoscopy video recordings	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD	Still TBD	Still TBD
DS1.4	Coronary computed	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
	tomography angiography (CCTA) variables, biometric data, medical history data			
DS2.1	Data related to hours spent by specialists.	Yes	Still TBD	Still TBD
DS2.2	Retrospective patient schedule data and precondition of the treatment	No	N/A	N/A
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	No	N/A	N/A
DS2.4	PROMs/PREMs	No	N/A	N/A
DS2.5	Prospective patient clinical data	No	N/A	N/A
DS2.6	Patient personal data (address, preferences, …)	No	N/A	N/A
DS2.7	Retrospective EHR	No	N/A	N/A
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	No	N/A	N/A

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices activation and connected loads	Data belongs to the owner of the location where the home automation devices are installed. The owner defines access to the data.	None	The owner defines access to the data.
DS3.2	IMU data captured by iPrognosis smartphone application	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	YES (unless otherwise decided due to IPRs)	Zenodo	Still TBD
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	YES (but considering the limitations related to the IPR/Consortium Agreement)	N/A	N/A

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	No	N/A	N/A
DS4.2	3D navigation system data (intracardiac signals, geometry)	No	N/A	Still TBD
DS5.1	Patient recordings and features extracted from interaction with patients	NO for recording, YES for features and patient is replaced by a random number during anonymization	EOSC, Zenodo	Still TBD, preferred open for research (e.g. CC-BY-NC)
DS5.2	Biometric data (e.g., blood pressure and heart rate)	YES for cohorts	EOSC, Zenodo	Still TBD, preferred open for research
DS5.3	Retrospective electronic health records	NO	N/A	N/A
DS5.4	PREMs related to clinical staff (depends on T1.4)	YES (but considering anonymization)	Zenodo	Still TBD, preferred open for research (e.g. CC-BY-NC)
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	YES (but considering anonymization)	Zenodo	Still TBD, preferred open for research (e.g. CC-BY-NC)
DS5.6	Datasets for facial expression and emotion recognition	YES	e.g., jaffe, FER-2013, MMI, Cohn-Kanade, RaFD, FERG, EMOTIC, Affect data,	Licenses are granted on individual requests, not handled by UM but the owners of data sets

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
			SemEval-2017 Task 4, DailyDialog, The MPLab GENKI Database, FFECTIVA-MIT Facial Expression Dataset (AM-FED), Grounded Emotions, Reldi, etc.	
DS5.7	Datasets for ASR and TTS in Slovenian	NO		Access can be granted on individual basis (bilateral agreements which may include charges)
DS5.8	Datasets for ASR and TTS in French	NO		Access can be granted on individual basis (bilateral agreements which may include charges)
DS5.9	EVA Corpus, data set of conversational expression	YES	Clarin.SI	CC-BY 4.0 License
DS5.10	Video recordings of third persons	NO	N/A	N/A
DS5.11	Patient's behavioural information	NO	N/A	N/A
DS5.12	Patient's facial information	NO	N/A	N/A
DS6.1	Patient's physical, medical and mental status.	NO	N/A	N/A

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
	Vital signs with sensors data.			
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Still TBD	Still TBD	Still TBD
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Dataset made available openly?	Which open repository (or other available medium)?	Licenses, access identification, restrictions
	intervention @ UZB			
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Still TBD	Still TBD	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	Still TBD	Still TBD	Still TBD
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Still TBD	Still TBD	Still TBD
DS8.1	Image, gene, phenotype and pathology data for glioma patients	No	Not relevant	Still TBD
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	Applicable for pilot specifics	Applicable for pilot specifics	Applicable for pilot specifics

A.7 Data interoperability

Table 14: Data interoperability.

Code	Type of data to be collected / name of the dataset	Metadata vocabularies, ontologies, standards and methodologies for data interoperability
DS1.1	Cardiac ultrasound video recordings	HL7
DS1.2	Capsule endoscopy video recordings	HL7
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Still TBD
DS2.1	Data related to hours spent by specialists.	Still TBD
DS2.2	Retrospective patient schedule data and precondition of the treatment	JSON
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	JSON
DS2.4	PROMs/PREMs	Still TBD
DS2.5	Prospective patient clinical data	HL7 FHIR
DS2.6	Patient personal data (address, preferences, …)	Still TBD
DS2.7	Retrospective EHR	JSON
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	N/A
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	N/A
DS3.2	IMU data captured by iPrognosis smartphone application	HL7

Code	Type of data to be collected / name of the dataset	Metadata vocabularies, ontologies, standards and methodologies for data interoperability
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	HL7
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	HL7
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	HL7
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	FHIR v4
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	N/A
DS4.2	3D navigation system data (intracardiac signals, geometry)	Still TBD
DS5.1	Patient recordings and features extracted from interaction with patients	FHIR v4
DS5.2	Biometric data (e.g., blood pressure and heart rate)	FHIR v4
DS5.3	Retrospective electronic health records	N/A
DS5.4	PREMs related to clinical staff (depends on T1.4)	FHIR v4
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	FHIR v4
DS5.6	Datasets for facial expression and emotion recognition	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	
DS5.8	Datasets for ASR and TTS in French	N/A
DS5.9	EVA Corpus, data set of conversational expression	N/A
DS5.10	Video recordings of third persons	N/A
DS5.11	Patient's behavioural information	FHIR v4
DS5.12	Patient's facial information	FHIR v4

Code	Type of data to be collected / name of the dataset	Metadata vocabularies, ontologies, standards and methodologies for data interoperability
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	N/A
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Still TBD
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	Still TBD
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Still TBD
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Still TBD
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	Applicable for pilot specifics

A.8 Data licensing, availability and usability

Table 15: Data licensing, availability and usability by third parties.

Code	Type of data to be collected / name of the dataset	Availability and usability of data by third parties. Licensing
DS1.1	Cardiac ultrasound video recordings	Still TBD
DS1.2	Capsule endoscopy video recordings	Still TBD
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Still TBD
DS2.1	Data related to hours spent by specialists.	Consortium agreement
DS2.2	Retrospective patient schedule data and precondition of the treatment	Consortium agreement
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	Consortium agreement
DS2.4	PROMs/PREMs	Internal use only
DS2.5	Prospective patient clinical data	Internal use only
DS2.6	Patient personal data (address, preferences, …)	Internal use only
DS2.7	Retrospective EHR	Consortium agreement
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction.	N/A
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Still TBD
DS3.2	IMU data captured by iPrognosis smartphone application	Still TBD
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Still TBD

Code	Type of data to be collected / name of the dataset	Availability and usability of data by third parties. Licensing
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Still TBD
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Still TBD
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Consortium agreement
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	N/A
DS4.2	3D navigation system data (intracardiac signals, geometry)	Still TBD
DS5.1	Patient recordings and features extracted from interaction with patients	N/A for videos, for features and cohorts Still TBD, CC BY-NC 4.0 is preferred
DS5.2	Biometric data (e.g., blood pressure and heart rate)	Still TBD, CC BY-NC 4.0 is preferred
DS5.3	Retrospective electronic health records	N/A
DS5.4	PREMs related to clinical staff **(**depends on T1.4**)**	Still TBD, CC BY-NC 4.0 is preferred
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	Still TBD, CC BY-NC 4.0 is preferred
DS5.6	Datasets for facial expression and emotion recognition	Licenses are granted on individual requests, not handled by UM but the owners of data sets
DS5.7	Datasets for ASR and TTS in Slovenian	Access can be granted on individual basis (bilateral agreements which may include charges)
DS5.8	Datasets for ASR and TTS in French	Access can be granted on individual basis (bilateral agreements which may include charges)
DS5.9	EVA Corpus, data set of conversational expression	CC-BY 4.0 License
DS5.10	Video recordings of third persons	N/A
DS5.11	Patient's behavioural information	N/A
DS5.12	Patient's facial information	N/A

Code	Type of data to be collected / name of the dataset	Availability and usability of data by third parties. Licensing
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	N/A
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Still TBD
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Still TBD
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	Still TBD
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Still TBD
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Still TBD
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	Applicable for pilot specifics

A.9 Data quality assurance

Table 16: Data quality assurance.

Pilot nr	Task nr	Type of analysis	Will you work according to specific protocol(s)? If yes, which one(s)?	Who will create the statistical analysis plan? (partner short name; person name; email)	How will data transformation & analysis be verified?
1	5.2	Still TBD	Still TBD	AUTH; person(s) Still TBD	Peer-review
2	3.2	Still TBD	Verification and validation plan	CHU de Liège, Patrick Duflot	System test, Unit test, Manual test
2	5.2	Still TBD	Prospective study protocol is under preparation	CHU de Liège, Patrick Duflot	As defined in clinical study protocol
2	5.3	As defined by PhE	As defined by PhE	As defined by PhE	As defined by PhE
3	3.3	Still TBD	Still TBD	Still TBD	Still TBD
4	3.5	Still TBD	Still TBD	INTRAS (person(s) Still TBD)	Still TBD
4	5.2	Still TBD	Still TBD	Still TBD	Still TBD
5	3.5	Protocols published for the pilot 5 Clinical studies	Verification and validation plan, Data Monitoring Plan	UM, GC, ITCL	Peer-review, Manual test
6	3.5	Protocols shared by PhE	Protocols shared by PhE	ITCL	Still TBD
7	3.6	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD

Pilot nr	Task nr	Type of analysis	Will you work according to specific protocol(s)? If yes, which one(s)?	Who will create the statistical analysis plan? (partner short name; person name; email)	How will data transformation & analysis be verified?
7	3.6	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD
7	3.6	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD
7	3.6	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD

Pilot nr	Task nr	Type of analysis	Will you work according to specific protocol(s)? If yes, which one(s)?	Who will create the statistical analysis plan? (partner short name; person name; email)	How will data transformation & analysis be verified?
7	3.6	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention@ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD
7	3.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention@ UZB	Retrospective study protocol (registry) is under preparation	Philips Image Guided Therapy Systems, Netherlands	Still TBD
7	3.6	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT of data.	Prospective study protocol (comparative observational study) is under preparation	Primary investigator UZB, VUB (Still TBD)	Comparative analysis of different key performance indicators and patient reported outcome and experience measures (PREMS and PROMS) evolution before (control

Pilot nr	Task nr	Type of analysis	Will you work according to specific protocol(s)? If yes, which one(s)?	Who will create the statistical analysis plan? (partner short name; person name; email)	How will data transformation & analysis be verified?
					period) and after implantation of AI smart cathlab prototype (study period)
8	N/A	Image, gene, phenotype and pathology data for glioma patients	To be determined in collaboration with UZ Brussels ICT	VUB; to be determined	Local testing, feedback from specialists, eventual peer review
PhE	5.3	Based on Quality control Standard Operating Procedure (SOP) followed by PhE for all projects/deliverables	Protocol & CRF has been prepared by PhE and will be shared with all pilots in order to follow the same	PhE project team / Eugena Stamuli, Declan O' Byrne, Magda Chatzikou	Still TBD

A.10 Data cleansing, transforming and analysing

Table 17: Data cleansing, transforming and analysing.

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
DS1.1	Cardiac ultrasound	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
	video recordings						
DS1.2	Capsule endoscopy video recordings	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS2.1	Data related to hours spent by specialists.	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS2.2	Retrospective patient schedule data and precondition of the treatment	Extraction and alignment from subsystems.	CHUL	Mapping to FHIR (responsible TMA)	None	In HosmartAI platform	
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	Extraction and alignment from subsystems.	CHUL	Mapping to FHIR (responsible TMA)	None	In HosmartAI platform	
DS2.4	PROMs/PREMs	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS2.5	Prospective patient clinical data	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
DS2.6	Patient personal data (address, preferences, …)	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS2.7	Retrospective EHR	Extraction and alignment from subsystems.	CHUL	Mapping to FHIR (responsible TMA)	Still TBD	Still TBD	
DS2.8	Data from radiotherapy services, infrastructure. Patient's satisfaction	N/A	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS3.1	Smart home data: Consumption/ production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Outlier detection, inference on missing data	VIMAR	Normaliza-tion	Still TBD	Still TBD	Still TBD
DS3.2	IMU data captured by iPrognosis smartphone application	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	N/A	N/A	N/A	N/A	N/A	N/A
DS4.2	3D navigation system data (intracardiac signals, geometry)	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS5.1	Patient recordings and features extracted from interaction with patients	Presence of face and audio; Data cleaning will be carried out by UM	UM	N/A	OpenCV and other tools developed by UM	At the edge, semi-automatic	Still TBD
DS5.2	Biometric data (e.g., blood	Data cleaning	UKCM	Still TBD	Still TBD	Still TBD	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
	pressure and heart rate)	will be performed by the data controller (UKCM)					
DS5.3	Retrospective electronic health records	N/A	N/A	N/A	N/A	N/A	N/A
DS5.4	PREMs related to clinical staff (depends on T1.4)	Data cleaning, such as removal of partial answers, duplication, etc. will be performed by the data controller (UKCM)	UM, UKCM	Still TBD	Still TBD	At the edge, semi-automatic	N/A
DS5.5	PREMs related patients (PAM, SUS-SI/TAM, UEQ)	Data cleaning will be performed by the data controller (UKCM)	UM, UKCM	Still TBD	Still TBD	At the edge, semi-automatic	Still TBD
DS5.6	Datasets for facial expression and emotion recognition	N/A	N/A	N/A	N/A	N/A	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	N/A	N/A	N/A	N/A	N/A	N/A
DS5.8	Datasets for ASR and TTS in French	N/A	N/A	N/A	N/A	N/A	N/A
DS5.9	EVA Corpus, data set of	N/A	N/A	N/A	N/A	N/A	N/A

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
	conversational expression						
DS5.10	Video recordings of third persons	N/A	N/A	N/A	N/A	N/A	N/A
DS5.11	Patient's behavioural information	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS5.12	Patient's facial information	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD	Still TBD
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data.	Gradior data cleaning will be performed by data owner (INTRAS) Rest of the data TBD	RAS will be responsible for data cleaning and preparation.	Still TBD	Excel/ SQLSERVER	Still TBD	Still TBD
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Data cleaning will be performed by the data owner (UZB)	UZB primary investigator will be responsible for data cleaning and preparation (Still TBD)	Manual removal of outlier patients (low image quality)	Excel	Philips Image Guided Therapy Systems, Netherlands	Still TBD
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Data cleaning will be performed by the data owner (UZB)	UZB primary investigator will be responsible for data cleaning and preparation (Still TBD)	Manual removal of outlier patients (low image quality)	Excel	Philips Image Guided Therapy Systems, Netherlands	Still TBD
DS7.3	Coronary angiogram imaging data of patients who undergone a	Data cleaning will be performed by the	UZB primary investigator will be responsible for data	Manual removal of outlier patients	Excel	Philips Image Guided Therapy Systems, Netherlands	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
	coronary angiogram/coronary intervention @ UZB	data owner (UZB)	cleaning and preparation (Still TBD)	(low image quality)			
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Data cleaning will be performed by the data owner (UZB)	UZB primary investigator will be responsible for data cleaning and preparation (Still TBD)	Manual removal of outlier patients (low physiology curves quality)	Coroventis software and Virtual stenting algorithm (VSA) for the interoperation of RFR/FFR pullback created by JF Argacha and Jean Decamp (I-depot number 123060 Date 16-04-2020)	Philips Image Guided Therapy Systems, Netherlands And JF Argacha, cardiology department, UZB, VUB, Brussel	Still TBD
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Data cleaning will be performed by the data owner (UZB)	UZB primary investigator will be responsible for data cleaning and preparation (Still TBD)	Manual removal of outlier patients (low image quality)	Optis and volcano software	Philips Image Guided Therapy Systems, Netherlands	Still TBD
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary	Data cleaning will be performed by the data owner (UZB)	UZB primary investigator will be responsible for data cleaning and preparation (Still TBD)	Manual removal of outlier patients (low image quality)	Philips and Heartflow software	Philips Image Guided Therapy Systems, Netherlands	Still TBD

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
	angiogram and/or a coronary intervention						
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Not applicable (prospective inclusion)	Not applicable (prospective inclusion)	Not applicable (prospective inclusion)	Not applicable (prospective inclusion)	Philips Image Guided Therapy Systems, Netherlands And Cardiology department UZB, VUB	Still TBD
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Connecting sample information per patient across databases, machine learning on data	To be hired	Integration, neural networks	Python, pytorch, sklearn, XNAT framework	VUB	Still TBD
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	Data cleaning will be performed by each data owner (pilot). PhE will perform the analysis on clean datasets	Still TBD by each pilot.	Economic & PRO/PREM analysis (economic evaluation, cost consequence analysis, cost-utility analysis, patients' quality of life measurem	Stata, Excel, maybe SAS if necessary	PhE	Following ISPOR guidelines for economic evaluation and Dolan's publication on EQ-5D analysis

Code	Type of data to be collected / name of the dataset	Type of data cleaning needed (e.g., correct data types, remove duplicates, add missing info…)	Person responsible for data cleaning (partner short name; person name; email)	Type of data transformation/analysis (e.g., normalization, discretization, …)	Software/tools used for cleaning, transform, and analyse	Where/by whom will the analysis be conducted?	Standards followed for code development / access and re-use
				ent, bootstrapping, regression, etc			

A.11 Ethical review

Table 18: Ethical review.

Code	Type of data to be collected / name of the dataset	Type of ethical review needed
DS1.1	Cardiac ultrasound video recordings	Clinical protocol, Ethical approval by Institutional Research Ethics Board
DS1.2	Capsule endoscopy video recordings	Clinical protocol, Ethical approval by Institutional Research Ethics Board
DS1.3	Cardiotocography variables and results, biometric data, medical history data	Clinical protocol and ethical approval from the hospital administration
DS1.4	Coronary computed tomography angiography (CCTA) variables, biometric data, medical history data	Clinical protocol and ethical approval from the hospital administration
DS2.1	Data related to hours spent by specialists.	N/A
DS2.2	Retrospective patient schedule data and precondition of the treatment	DPO approval
DS2.3	Data linked to radiotherapy machines (tumours treatment indication, maintenance, building location)	N/A
DS2.4	PROMs/PREMs	Ethical approval
DS2.5	Prospective patient clinical data	Ethical approval
DS2.6	Patient personal data (address, preferences, …)	Ethical approval
DS2.7	Retrospective EHR	DPO approval
DS2.8	Data from radiotherapy services, infrastructure.	N/A

Code	Type of data to be collected / name of the dataset	Type of ethical review needed
	Patient's satisfaction	
DS3.1	Smart home data: Consumption/production of instantaneous electricity and consumption logs, Human presence/access control, Devices' activation and connected loads	Clinical protocol, Ethical approval by the national review board
DS3.2	IMU data captured by iPrognosis smartphone application	Clinical protocol, Ethical approval by Research Ethics Committee
DS3.3	Voice-related time and spectral features captured by iPrognosis smartphone application	Clinical protocol, Ethical approval by Research Ethics Committee
DS3.4	Key taps press and release timestamps captured by iPrognosis smartphone virtual keyboard	Clinical protocol, Ethical approval by Research Ethics Committee
DS3.5	Joint coordinates of 3D skeleton data captured by the iPrognosis iMAT application	Clinical protocol, Ethical approval by Research Ethics Committee
DS3.6	Results of rehabilitation sessions with Gradior and care plans (editor and patient data)	Clinical protocol, Ethical approval by Research Ethics Committee
DS4.1	Demographic patient data (age, gender, atrial size, etc.)	Still TBD
DS4.2	3D navigation system data (intracardiac signals, geometry)	Still TBD
DS5.1	Audio-Visual recordings of patients	Clinical protocol, Ethical approval by the hospital's Ethics Committee
DS5.2	Biometric data (e.g., blood pressure and heart rate)	Clinical protocol, Ethical approval by the hospital's Ethics Committee
DS5.3	Retrospective electronic health records	Clinical protocol, Ethical approval by the hospital's Ethics Committee
DS5.4	PREMs related to clinical staff (depends on T1.4)	Clinical protocol, Ethical approval by the hospital's Ethics Committee

Code	Type of data to be collected / name of the dataset	Type of ethical review needed
DS5.5	PREMs related patients (e.g., PAM, SUS-SI/TAM, UEQ) and PROs	Clinical protocol, Ethical approval by the hospital's Ethics Committee
DS5.6	Datasets for facial expression and emotion recognition	N/A
DS5.7	Datasets for ASR and TTS in Slovenian	N/A
DS5.8	Datasets for ASR and TTS in French	N/A
DS5.9	EVA Corpus, data set of conversational expression	N/A
DS5.10	Video recordings of third persons	N/A
DS5.11	Patient's behavioural information	N/A
DS5.12	Patient's facial information	N/A
DS6.1	Patient's physical, medical and mental status. Vital signs with sensors data	N/A
DS7.1	Administrative data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Clinical protocol (patient registry). Ethical approval by the hospital review board will be asked.
DS7.2	Clinical data of patients scheduled for a coronary angiogram/coronary intervention @ UZB	Clinical protocol (patient registry). Ethical approval by the hospital review board will be asked.
DS7.3	Coronary angiogram imaging data of patients who undergone a coronary angiogram/coronary intervention @ UZB	Clinical protocol (patient registry). Ethical approval by the hospital review board will be asked.
DS7.4	Coronary physiology data of patients evaluated by a resting index measure (iFR/RFR) or a hyperemic index (FFR) either during a manual or a motorized wire pullback and performed before and after a coronary intervention @ UZB	Clinical protocol (patient registry). Ethical approval by the hospital review board will be asked.
DS7.5	Intravascular imaging data of patient evaluated by either OCT or IVUS technique before and after a coronary intervention	Clinical protocol (patient registry). Ethical approval by the hospital review board will be asked.

Code	Type of data to be collected / name of the dataset	Type of ethical review needed
DS7.6	Coronary CT data including FFRCT computation of patient referred for an invasive coronary angiogram and/or a coronary intervention	N/A
DS7.7	Full prospective UZB data set of administrative, clinical, coronary angiogram, coronary physiology, intravascular imaging and coronary CT data.	Clinical protocol (prospective comparative study). Ethical approval by the hospital review board will be asked.
DS8.1	Image, gene, phenotype and pathology data for glioma patients	Already performed and approved.
DS.O.1	All pilot data with Common KPIs (Economic & PROMs/PREMs)	The economic & PROMs/PREMs data will be included in the protocol.