Improving cancer diagnosis and prediction with AI and big data

A Multimodal AI-based Toolbox and an Interoperable Health Imaging Repository for the Empowerment of Imaging Analysis related to the Diagnosis, Prediction and Follow-up of Cancer

Deliverable 7.1

Initial Data Management Plan

WP7 – Legal and Ethics Management

31-03-2021

Revision 1.0

Status: Final

Grant Agreement n 952179

European Commission

DOCUMENT CONTROL	
Project reference	Grant Agreement number: 952179
Document name	D.7.1 Initial Data Management Plan
Work Package	WP7
Work Package Title	Legal and Ethics Management
Dissemination level	PU
Revision	1.0
Status	Final
Reviewers	Shereen Nabhani (KU), Ioanna Chouvarda (AUTH)
Beneficiary(ies)	Timelex

Dissemination level:

PU = Public, for wide dissemination (public deliverables shall be of a professional standard in a form suitable for print or electronic publication) or CO = Confidential, limited to project participants and European Commission.

AUTHORS		
	Name	**Organisation**
Document leader	Jos Dumortier, Magdalena Kogut – Czarkowska	TLX
Participants	All Partners	KU, BSC, ICCS, UNI, WR, MDT, FTSS

REVISION HISTORY				
Revision	**Date**	**Author**	**Organisation**	**Description**
0.1	3/2/2021	Magdalena Kogut - Czarkowska	TLX	Structure and scoping
0.1.1	26/2/2021	Magdalena Kogut - Czarkowska	TLX	Initial draft for Partners' input
0.1.2	10/3/2021	Magdalena Kogut - Czarkowska	TLX	Consolidated Partners' input

REVISION HISTORY				
Revision	**Date**	**Author**	**Organisation**	**Description**
0.1.3	16/3/2021	Magdalena Kogut - Czarkowska	TLX	Clean version for internal review
0.1.4	19/3/2021	Jos Dumortier	TLX	Internal review
0.2	21/3/2021	Magdalena Kogut - Czarkowska	TLX	Draft for peer review
0.3	30/3/2021	Magdalena Kogut - Czarkowska	TLX	Draft incorporating input from peer – reviewers
1.0	31/3/2021	Magdalena Kogut - Czarkowska	TLX	Final version ready for submission

Disclaimer and statement of originality

The content of this deliverable represents the views of the authors only and is their sole responsibility; it cannot be considered to reflect the views of the European Commission and/or the Consumers, Health, Agriculture and Food Executive Agency or any other body of the European Union. The European Commission and the Agency do not accept any responsibility for use of its contents.

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Table of Contents

Terms and Abbreviations

Term	Description
Project	Horizon 2020 Research and Innovation action called "A Multimodal AI-based Toolbox and an Interoperable Health Imaging Repository for the Empowerment of Imaging Analysis related to the Diagnosis, Prediction and Follow-up of Cancer (Grant Agreement No: 952179)
Data Provider	The following Consortium Partners: AUTH, HCS, UoA, UNITOV, DISBA, GOC, UNS, VIS, IDIBAPS
Federated Repository	Pan-European Repository of Health Images
AI Toolbox	AI solutions to improve cancer detection systems and enhance the clinical workflow
Temporary Infrastructure	IT infrastructure stack hosted by FTSS
Consortium Partners or Partners	Partners to the Grant Agreement n 952179

Abbreviation	Description
EC	European Commission
TCC	Technical and Clinical Committee
WP	Work Package
DMP	Data Management Plan
HCP	Heath Care Professional

1 Executive Summary

Data Management Plans (DMPs) are considered to be a key element to sound data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 Project.

The goal of this document is to set the initial DMP for the INCISIVE Project. It contains guidelines that will be used by the INCISIVE Consortium Partners with regards to all the data that will be generated by the Project.

This DMP is based on the DMP template provided by the European Commission (EC)[1] and follows the "Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020"[2] . Furthermore, in accordance with the requirements stated in the description of action covers the following aspects:

- Description of the data to be collected or created during research and solution deployment and piloting, including metadata.
- Standards and methodologies for data collection and management and quality assurance measures.
- Plans for data sharing and access.
- Copyrights and intellectual property of data.
- Data storage and back-up measures.
- Data management roles and responsibilities.

The present document constitutes the first version of the DMP and will be submitted within the first six months of the Project.

The DMP will not be a fixed document, but it will evolve and will gain more precision and substance during Project implementation. More detailed versions of the DMP will be delivered at later stages of the Project and will be concluded with the final Deliverable D7.5 Final Data Management Report.

The key factors which determine the increasing sophistication and level of detail of the DMP are (i) the stages of design and evolution of the AI tool solutions aimed to improve cancer detection systems and enhance the clinical workflow ("**AI Toolbox**") throughout the full cycle of their

[1] See European Commission, Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 26 July 2016.

[2] European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.

development and (ii) preparation and deployment of the Pan-European Repository of Health Images ("**Federated Repository**").

In the context of the design of the AI Toolbox, the methodology proposed by the Consortium Partners includes: (i) defining user requirements and initial collection of existing medical data for the purpose of designing 1st prototype of the AI solution and automated machine learning based annotation mechanism, (ii) testing of the prototype on additional data sets during pre-clinical trials and observational studies and preparing its improved model (2nd prototype) (iii) completion of the final prototype. This will entail collection of medical images and health data, such as retrospective and prospective training data and collecting surveys from patients and healthcare professionals (HCP) to gauge user requirements for the functionality of the AI Toolbox.

As to the Pan-European Repository of Health Images, data management constitutes the heart of this part of the Project, as the goal of the Federated Repository is to enable secure donation and sharing of data in compliance with ethical, legal and privacy demands, increasing accessibility to datasets for various stakeholders and enabling experimentation of AI-based solutions. The architecture of the Federated Repository will be defined during the Project.

As the Project is still in its initial months, in the present document we focus on the description of management of the data which is already being gathered by the Consortium Partners or will be produced/used soon, i.e.:

- Medical images and health data, in particular retrospective training data and
- INCISIVE needs assessment study data (WP2).

Those are the sets of data well defined at this point. For reasons of completeness, we list other datasets, which were indicated by the Consortium Partners, but we are not in a position to provide full answers from a DMP point view as the treatment of this future data will be part of the research work of the Project. For this future data we provide basic principles of management and will be able to make more detailed discussion in future updates of the DMP. We also indicate how the rules of data management will be incorporated in the future Federated Repository.

Hence, information of the future sets and design of the Federated Repository will be made available on a finer level of granularity through updates of the DMP as the implementation of the Project progresses and when significant changes occur, such as (but not limited to) the definition of new data protocols during the next stages of the development, progression of the tasks, changes in consortium policies, changes in consortium position and external factors (e.g. new Data Governance Act regulation).

2 Introduction

2.1 Purpose and scope

Purpose of this DMP is to:

- Create a document, which explains the management of data collected during the Project.
- Support the data management life cycle for all data that will be collected, processed or generated by the Project.
- Provide an analysis of the main elements of the data management policy, which will be used by the Partners regarding all the datasets which will be generated by the Project.
- Provide details and guarantee about the preservation of the data collected during the Project, as well as any results derived from the associated research.
- Provide details on how the INCISIVE consortium plans to address the ethical issues related to data, which will be collected during the Project timeframe.

The DMP is not a fixed document, but it will evolve during the lifetime of the Project.

2.2 Document structure

The rest of the document is structured as follows:

- Section 3 provides a brief description of data sets which will be collected during the INCISIVE Project, explains the procedures used to collect or create them, as well as standards and methodologies for data collection and management and quality assurance measures.
- Section 4 describes plans for data sharing and access in compliance with the FAIR principles.
- Section 5 deals with allocation of resources, data management roles and responsibilities.
- Section 6 discusses the security of the collected data, including data storage and back-up measures.
- Section 7 presents ethical issues, confidentiality and intellectual property of data.
- Section 8 contains conclusions and further plans for the updating of the DMP.

2.3 Relation with other deliverables

This deliverable is closely related to the following deliverable(s):

- D8.1: Innovation Strategy - First Version
- D8.3: Innovation Strategy - Second Version
- D8.6: Innovation Strategy - Final Version

The innovation strategies will contain identification of Partners' IP assets, including data, and an ownership proposition based on feedback from the consortium and the Project's IP mapping. It will also state the current or the suggested type of protection, and conditions of use for the Project's IP assets, including data.

- D7.3: Data Donation Legal Framework

Data donation legal framework will define the rules and standards regarding the data donation in the healthcare sector and form guidelines and terms of such donorship, which will be required to put forth the rules of sharing of data in the Federated Repository.

- D7.4: IPR Management Report

IPR management report, which will be closely related to innovation strategy, will document IPR assignment of the Project outcomes.

The deliverable also supports the activities within other work packages and tasks.

3 Data summary

3.1 Research data

The notion of “research data” refers to “information, in particular facts or numbers, collected to be examined and considered as a basis for reasoning, discussion or calculation”.[3] Research data covers a broad range of types of information, and digital data can be structured and stored in a variety of file formats. Examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images.

We note that properly managing data (and records) does not necessarily equate to sharing or publishing that data. Some kinds of data may not be sharable due to the nature of the records themselves, or to ethical and privacy concerns. This refers to, for example:

- Preliminary analyses
- Drafts of scientific papers
- Plans for future research
- Peer reviews
- Communications with colleagues

Research data which cannot be shared may also include trade secrets, commercial information, materials necessary to be held confidential by a researcher until they are published or similar information, which is protected under law.

3.2 Collection purposes.

While the detailed purposes of the data collection per each data set are outlined below, it may be summarized that the research data is collected and processed during the Project for the following purposes and in relation to the following Project objectives:

- Development of novel machine learning algorithms trained to support various categories of healthcare providers in dealing with the complexity of cancer imaging data, concretely in breast, lung, colorectal and prostate cancer.
- Labelling and annotation of cancer imaging data.
- Availability and sharing of data that can be used for training and validating AI tools for improved imaging methods.

In this regard, INCISIVE will develop and validate:

[3] European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.

- An AI-based toolbox to improve the accuracy, specificity, sensitivity, interpretability and cost-effectiveness of imaging methods for cancer.
- An ML-based automatic annotation system to produce data for the training of algorithms in Machine Learning (ML) research.
- A Pan-European Repository of Health Images that allows the donation and sharing of data (in compliance with legal, ethical, privacy and security obligations) for the experimentation in AI-based solutions.

The main results expected from INCISIVE's research work are:

- The INCISIVE AI-driven models enhancing image processing and data analysis focusing on improving sensitivity and specificity in diagnosis and statistical assessment of cancer.
- The INCISIVE Pan-European Repository of Health Images that will enable the secure access and sharing of data and ultimately allow the large-scale adoption of such solutions in cancer diagnosis and follow-up.
- The INCISIVE platform which through its HPC and HDPA-as-a-service, will provide secure and cost-effective performance of computationally intensive processing, without the need for maintaining expensive equipment.
- The INCISIVE user services and reporting tools in the form of intuitive and highly interactive visualizations, addressing the needs of stakeholders visualizing the analysis results along with corresponding reasoning, enabling the accurate detection, prediction and follow-up of cancer, and allowing decisions that are better informed.
- The INCISIVE anonymization mechanism aiming to enable legal/ethical sharing and processing of medical data.
- A pool of scientific publications in high ranked conferences and high impact journals.

3.3 Methodology of work

The specific data sets for the Project need to be identified and described with the contribution of all Project Partners. For this reason, all Project Partners were asked to describe the specific data sets that will be processed during the Project. Accordingly, early in the Project a table with the following questions was circulated to be filled by the WP and Task leaders, further complimented with input from other Partners.

1. OUTPUT DATA (NEW DATA)
    a) What new data will you gather or produce in this task and how?
    b) What is the purpose of the collection/generation of data in relation to the task?
    c) What is the expected size of dataset?

d) In what manner and format will the data be collected and kept?

e) What transformations will the data undergo?

f) What metadata will be created and used? Are there any standards applicable?

g) Will the data be commercially sensitive or otherwise confidential? Will there be any planned terms and conditions of its use?

2. INPUT DATA (RE-USE OF EXISTING DATA)

a) What existing data will you re-use for this task and for what purpose?

b) What is the source of existing data?

c) What transformations will the data undergo?

d) What is the expected size of existing data set?

e) In what manner and format will the data be kept and used?

f) What metadata will be created and used? Are there any standards applicable?

g) Is the data commercially sensitive or otherwise confidential? Are there any applicable terms and conditions of its use?

3. SOFTWARE TOOLS & STORAGE

a) With what IT tools will the data in this task be processed?

b) Where will the data be stored and backed-up?

c) How will the data be secured?

4. DATA SHARING

a) Will the data be shared during the Project? If yes, how and with whom?

b) Will the data be shared after the Project? If yes, how and with whom?

5. PRIVACY/ETHICS

a) Will the data – either new or existing - include any personal data?

b) If yes, will data subject obtain information about data processing and consent to it?

c) Will personal data be anonymized/pseudonymized?

d) Is there any ethics approval required?

The responses provided in the first months of the Project have been attached as **Annex 1**.

The DMP questionnaire is intended as a living document and more detail will be added as the Project progresses. All Partners commit to continuously keep track of the specific data sets processed under the tasks they are involved in and to report them internally by updating the DMP

questionnaire for each task of the Project. Partners will be periodically reminded to update their responses.

3.4 Data sets and data format.

Based on the provided input, at this early stage of the Project, the following main sets of research data were identified:

1) Medical images and health data

The following Partners: AUTH, HCS, UoA, UNITOV, DISBA, GOC, UNS, VIS, IDIBAPS ("**Data Providers**") will provide de-identified health and medical images data – retrospective training data - which will be used for purposes of subsequent research tasks within WP3, WP4 and WP5. Retrospective training data and prospective training data (described in more detail below) will be the initial subsets of medical images and health data which will be collected within the Project.

Retrospective training data will be extracted by each Data Provider from their hospital systems and uploaded to Temporary Infrastructure (as further described in Section 6 below). This data collection is planned to start in May 2021, or earlier if all legal requirements are fulfilled. The data will comprise a wide variety of patient data and existing cancer images (CT, MRI, PET/CT, Ultrasounds, X-Rays, Mammographs etc). This data will be provided in. dcm and .jpg formats.

In addition to the medical images, data, additional demographics and general information about the patient will be provided in .txt format:

- year of birth, age of diagnosis, gender, history of family cancer (without personal information regarding family members), medication history, symptoms, etc.,
- histology and detected markers (directly or indirectly related), laboratory tests,
- diagnostic material (e.g., staging information, tumour location etc.) will also be collected for all cancer types at distinct time points (the timepoints considered are: baseline, after 1st treatment, 1st follow up, 2nd follow up).

The retrospective training data will enable the excessive training of the foreseen AI Toolbox. The exact scope of personal data processed is defined in a protocol. The retrospective data will be combined with prospective training data.

Prospective training data will involve additional collection of data from current or future patients who are receiving diagnoses and/or treatment from the Data Providers. Each of the Data Providers will identify current patients or new patients and obtain their consent for including their data in the study. The protocol for this data collection will be discussed and decided by all Data Providers in the course of the Project and will be the next step after concluding with the existing retrospective data. This part will start as soon as the retrospective data has been ethically cleared and collection has almost been completed. This data will be added to retrospective data, to allow further training and validation of the AI Toolbox.

As further noted below, all medical images and health data will be de-identified or anonymized prior to sharing and the algorithms will not take a patient's name or other identifying metadata into consideration to analyse their medical images.

The retrospective and prospective data sources will be augmented with open-source data. These open-source data will be selected based on their relevance according to the Project objectives, their availability and compatibility with the retrospective data. They will be digested and formatted based on the needs identified by the consortium. The proper acknowledgment of these open sources and of the work of others will be guaranteed by applying appropriate citation and quotation methods.

These medical images and health data (encompassing retrospective training data, prospective training data and open-source data, and patient data from study pilots, which will be collected within later stage of the Project) will be used to accomplish various goals of the Project, in particular to train the AI Toolbox, to develop final Federated Repository and develop donorship mechanism. The consortium will also consider possible terms of making the medical images and health data available in the Federated Repository (as outlined further below).

2) INCISIVE needs assessment study data (WP2)

These data sets collected within WP 2, T2.1. consist of feedback obtained in response to surveys and interviews with health care professionals and patients within the following studies:

- For HCPs within the INCISIVE consortium: "Identification of Care pathway for cancer diagnosis, recurrence and treatment response and post-treatment care". This involves email semi-structured interviews with HCPs within the Project consortium. The interview schedule consists of 28 questions exploring the care pathway in each country. After collection, data will be entered into Microsoft excel spreadsheet to allow content examination and comparative analysis of the data.
- For HCPs involved in cancer care (i.e.: primary users of INCISIVE): "Identification of healthcare professionals' perceptions and experiences with cancer care and requirements for INCISIVE". This involves surveying HCPs using online data collection method. The survey consists of 58 questions divided into 6 sections.
- For Patients: "Perceptions and experiences of cancer patients regarding existing cancer care pathways across Europe". The interview consists of 15 open ended questions. Interviews were audio-recorded and then transcribed verbatim into text.

3) Other

Apart from the principal categories mentioned above, the following main sets of data were identified as to be collected, processed and/or generated in the course of the Project, during its later stages:

- Outcomes produced during UX design workshops with participation of INCISIVE stakeholders (T2.2). Design thinking method and Delphi approach will be used as methodological approaches to better understand and prioritise users' requirements.
- Experimental data deriving from the evaluation of the impact of the Trusted Execution Environment technology on the overall system performance, and administrative metadata related to user management will be collected to optimize the performance of the user-centre design (T3.3).
- System telemetry from INCISIVE applications running on our HPC testbed systems and supercomputer (T3.4).
- AI algorithms and analysis of their performance for semantic segmentation of medical images, disease detection and classification (e.g. tumour staging), as well as disease progression estimation. Both image processing and AI algorithms will be applied on anonymised patient data with the goal of improving the diagnosis on cancer diseases. Their performance will be evaluated on each case of cancer. Estimation models along with the aforementioned algorithms will be used to estimate the progress of the disease for each patient (various tasks in WP4).
- ML performance data - accuracy and precision evolution during ML processes (T4.8).
- Analytics on AI models' accuracy, sensitivity, specificity, efficiency and performance, validation results (T6.2).
- Patient data generated from the observational and interventional studies across the validation sites (T6.4). At each of the 8 pilot sites, the INCISIVE platform will be used to store and evaluate the data from a number of patients. The data will include relevant health data such as clinical data, imaging data, histopathology data, aligned with INCISIVE pilots' protocol definition for each cancer type (T2.6).
- Additional open-source and proprietary market intelligence related data for performing a thorough AI for Health Imaging Market Analysis; existing data and reports in Partners' repositories may also be used or reused (T8.1).
- Interviews and/or surveys aimed to validate and improve the INCISIVE business models as well as to establish direct contacts and relationships that may result in a more pro-active adoption of the Project's value propositions by target groups. Subjects will be stakeholders that have been identified as potential early adopters of the INCISIVE services and members of the advisory board. All subjects will be compiled in a respective contact database (T8.2).

- Stakeholders contacts, either from the subscriptions to newsletter or from all Partners to disseminate the state of the Project and inform about news related to the Project or events (T9.1 and 9.2).
- Project deliverables and management reports (public or confidential) as well as publications for scientific dissemination and other more general communication material will be generated by the Project Partners.

The specific data formats which have been identified so far are described in **Annex 1** in relation to each data set. Going further, the description of each of the data sets will be made more granular and detailed as the Project progresses, so that the specific information about each data sets for INCISIVE is provided with the contribution of all Project Partners.

3.5 Re-use any existing data and origin of the data

The Project will rely on existing data sets as well as produce new, original data.

1) Medical images and health data

The Project will rely on existing data sources for the initial development of the AI Toolbox. This existing data will be retrospective training data, i.e., de-identified previously collected patient data obtained in clinical context. The source of this data will be hospital systems of the Data Providers. Prior to sharing, this data will be de-identified according to agreed protocol and provided to Temporary Infrastructure. Ethical requirements which the consortium will follow with respect to use of this data are described below in Section 7.1.

Furthermore, this retrospective training data will be augmented with open-source data, which is another existing data source. List of the kind, source and number of open data sources to be used in the Project will be defined in the course of the Project tasks. Open-source data will not include any personal data.

2) INCISIVE needs assessment study data (WP2)

For studies regarding the better understanding of INCISIVE user needs and experiences, only original data will be collected. Neither the qualitative, nor the quantitative studies require any use of pre-existing or secondary data, provided that a theoretical background study based on the relevant existing literature is required to develop the Project studies.

3) Other

Also, other data sets described above in Section 3.4 will consist of original unpublished data, except where clearly indicated otherwise. Naturally, the data collected or developed during earlier tasks will fuel the progress of subsequent tasks. The original data will be collected by the Project Partners.

3.6 Expected size of the data

The total size of the research data is difficult to estimate at this point in the Project, however it will be substantial. Based on the responses collected so far, we identified the size of the following data sets.

1) Medical images and health data

The size of retrospective training data is estimated at approx. 2 PB of data provided by 9 Data Providers. After the initial development stage, the AI Toolbox will be investigated in four validation studies for Breast, Prostate, Colorectal and Lung Cancer, taking place in 8 pilot sites, from 5 countries (Cyprus, Greece, Italy, Serbia and Spain), with participation of estimated number of 2,550 patients and a total duration of 1,5 year. These studies will generate further research data.

Last, but not least, the ultimate objective of INCISIVE is to create the Federated Repository, which will store digital images of various cancer types. The estimated size of the data in the Federated Repository will be decided in the next Project stages.

2) INCISIVE needs assessment study data (WP2)

The size of the data that could be generated by studies regarding the better understanding of INCISIVE user needs and experiences will encompass:

- For HCPs (first round of interviews): responses from a sample of 7-10 oncology specialized HCPs within the Project consortium.
- For HCPs (surveys): to date 96 responses for the online survey have been collected. For Patients: 4 to 8 cancer survivors from each country representing the various available tumour types (35 - 40 participants in total).

Approximate data size is 1-2 GB.

3) Other

Further data will be generated during user requirements definition and system design workshops. Its estimated size is 1-2 GB. Also, the data obtained from performing an AI for Health Imaging Market Analysis, coming mainly from documents and reports, is expected not to exceed a maximum of 250 MBs.

Apart from the “raw” heath data, user interviews, and reports INCISIVE Project will also process metadata relating to the research studies underpinning these data and data relating to knowledge and training materials. In particular, the technical tasks related to development of the AI Tool, its training and transformations of existing medical images will also generate additional data. The size of this data is unknown at this point, and depends of various factors, such as medical image dataset volume.

3.7 Data utility, standards and quality assurance

3.7.1 Data utility

1) Medical images and health data

As its main goal, the Project aims to deliver a standalone Federated Repository, including mainly, but not only, medical images, build in accordance with FAIR principles, integrating data level along with functionalities for data sharing and identity management.

The repository will be built upon a federated storage approach and a set of standardized open APIs that will enable the linking of various local data sources, the interaction with the users, the communication with processing infrastructures and the sharing of data. Its stand-alone nature and interoperability features will enable its connection with third-party trusted AI providers, as well as with established healthcare systems (e.g., PIMED in Catalonia etc.), contributing to its future sustainability.

Therefore, the results of the Project research activities and Federated Repository established as a result of the Project will primarily be of benefit to academic researchers, clinicians and industrial Partners. Nevertheless, from a more holistic point of view, the data processing and analysis performed by means of the Federated Repository will contribute to the well-being of the society, through providing means to more precisely detect and treat cancer and develop new therapeutic approaches for diagnosis and prognosis.

2) INCISIVE needs assessment study data (WP2)

The studies conducted in WP2 will inform the design, development and implementation of the INCISIVE platform and will aim to benefit all stakeholders mentioned above, as they will have a reliable source of information, new insights and methods. In particular, they may be useful for AI researchers, machine learning and AI solution developers and decision-makers in the healthcare, AI and innovations sector.

3) Other

The consortium will examine whether any other data results generated as a result of the Project research activities could be relevant by made identified Project stakeholders, i.e. all actors involved with AI and digital images matters in the healthcare sector, such as healthcare professionals, relevant industrial Partners, researchers and developers, policy- and other decision-makers, the European societies and respective and the academia and made available to them via the online portal for the purpose of making them reusable. For example, although AI training materials will be specific to INCISIVE, however, their strategy and approach can be extrapolated to other AI solutions. As such they may be a point of reference for other healthcare stakeholders, in particular for HCP which use AI solutions.

3.7.2 Standards

INCISIVE will make use of existing standards where applicable and make pre-standardization work suggestions for contributing to existing efforts especially focusing on related ontologies, vocabulary, image labelling, annotation and anonymization. The Partners were asked to identify the standards for each dataset in the spreadsheet. While it is still early in the Project, the standards indicated below were identified:

1) Medical images and health data
    - DICOM, in particular DICOM PS3.15 2021a - Security and System Management Profiles;
    - HL7 FHIR for interoperability and data exchange issues in relation to medical images data.

2) INCISIVE needs assessment study data (WP2)
    - N/A
3) Other
    - Risk management in medical devices (ISO 14971) as well as IEC 62304 (Medical device software — Software life cycle processes);
    - SNOMED, LOINC or other suitable terminologies will be used for data formalization;
    - Software quality standards for the developed software and other standards described in the D1.1 Project Management Handbook for Project deliverables.

Additionally, the Ethics guidelines for trustworthy AI guidelines[4] will be considered, as well as IEEE standards and ongoing standardisation activities.[5]

Furthermore, within dedicated Project task, the consortium will undertake the alignment of INCISIVE with identified and corresponding standards by providing standardization guidelines and actions derived from the analysis of data provided by Data Providers. One of the identified challenges is the lack of proper standards and APIs for interoperability with existing hospital information systems and Electronic Health Records (EHRs), which hamper integration of AI solutions in cancer imaging into clinical practice. Deliverable D3.4 Standardization Suggestions will be produced which will contribute to existing standards and stimulate creation of new ones. An area also targeted in the Project will be “legal standardization” of the donorship data sharing process and its underlying interoperability layer, which will be addressed in Deliverable D7.3 Data Donation Legal Framework.

[4] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai
[5] https://standards.ieee.org/initiatives/artificial-intelligence-systems/standards.html, including https://standards.ieee.org/project/2841.html,https://standards.ieee.org/project/2830.html, https://standards.ieee.org/project/2802.html and https://standards.ieee.org/project/2801.html

3.7.3 Quality assurance

Quality Assurance will take place during the Project according to the procedures set out in the context of T1.3 addressing quality assurance.

The quality assurance procedures to be applied on the data (re)used and/or generated in each of the INCISIVE work packages are defined in D1.2 - Quality Assurance and Risk Management plan and are, therefore, not repeated here in detail. The below are indicative examples of different quality assurance processes followed in INCISIVE for addressing data quality challenges for specific major project datasets and not a full list of all data quality assurance processes in the Project WPs.

1) Medical images and health data

The quality of the retrospective data will be guaranteed by the appropriate approvals of ethical boards collected within WP7, which outline how the data may be extracted from the original datasets. Furthermore, the Consortium Partners will agree on the data anonymization and annotation protocols, prior to using the data for training purposes. In particular, the quality of health data, medical images and their metadata (annotations) in the context of their utility for the development of the AI toolbox shall be ensured through a series of workshops organized between the technical Partners and the Data Providers. The aim of those workshops is to reach a common agreement on the scope of data required by AI toolbox to solve the challenges described in the action. The annotation protocols shall ensure that the medical information that are required for training the algorithms are extracted by medical professionals and are available to the technical Partners to develop their models accordingly.

Within WP5 the relevant Partners will contribute to the quality assurance of the INCISIVE data in the Federated Repository through the implementation of automated data curation, annotation and anonymisation mechanisms.

2) INCISIVE needs assessment study data (WP2)

The quality of the WP2 data that is being collected from health care professionals and patients is addressed through the methodology of work for data collection defined before starting the collection of data.

3) Other

General quality assurance processes will be followed. These quality assurance procedures are being further detailed and specified within each work package as the Project progresses.

4 Findable, Accessible, Interoperable and Re-usable data (FAIR data)

4.1 Making data findable, including provisions for metadata

INCISIVE Project attaches great importance to making its research data findable, discoverable and identifiable. The DMP defines what documentation and metadata will accompany the data.

"Metadata" is structured information describing the characteristics of a resource. For example, the dates associated with a dataset or the title and author of a book. Metadata supports discovery, re-use and long-term preservation of resources. Metadata needs to vary across scientific fields, but typically cover the following:

- Descriptive metadata, such as title, abstract, author, and keywords;
- Administrative metadata which are used to provide information to help manage a source, such as when and how it was created, file type and other technical information, and who can access it;
- Archive terms and access policies.

A metadata record consists of a set of predefined elements that define specific attributes of a resource. Each element can have one or more values; for example, a dataset may have multiple creators or more keywords may be added to a particular image to enable its finding. Documenting data enables other researchers to discover the data. Metadata about the nature of the files is also critical to the proper management of digital resources over time.

In this section of the document, we provide an outline regarding the application of FAIR principles to research data identified at this stage of the Project. We also provide a description on how those principles will be enshrined in the Federated Repository, the ultimate product of the Project that will make the results of this action available to public.

1) Medical images and health data

At the current stage of the Project, where retrospective data will be shared by the Data Providers, this data will come with their own metadata stemming from each of those providers. Thus, the Partners will agree on specific issues regarding providing adequate metadata within the dataset (e.g. field labels or column headings) in order to be easy to interpret the data. Such agreement is already being developed for retrospective training data, during the definition of the retrospective study protocol. The metadata for the annotation of the prospective data will be defined from the beginning, so it will be a lot easier to harmonize.

The challenge which INCISIVE will try to solve is to automatically or semi-automatically harmonize the metadata so that the reusability, findability and interoperability of this data is improved. Thus, within T5.2 (Automated data curation, annotation and anonymization mechanisms) of the Project, the responsible Partners will undertake annotation both of the raw data, as well as of the produced results, in order to enhance standardization, explainability and findability of this data. The main aim of this processes is to be performed in an automated and standardized way, to the

extent possible, minimizing the human intervention. The annotation and labelling of images will consist in designing methods of automatic addition of information to the provided data (ex. tumour location), which will be available in the Federated Repository, to allow this data to be used as training data for machine learning research. The outputs of these tasks will frame an intermediate layer between the data available and the analysis and reuse of this data for training and validation of AI.

2) INCISIVE needs assessment study data (WP2)

For study results, the metadata will consist of a report describing the procedures for data collection including: the number of participants for survey and interviews, the data collection period, including the dates for conducting the interviews.

3) Other

Within other datasets identified at this point, it was indicated in particular that:

- for system telemetry data from INCISIVE applications, the collected metadata will consist of application/execution ID, system configuration and additional execution environment data;
- for algorithms performance data, metadata will consist in application/execution ID, ML hyperparameters and system tuning.

For other data, metadata records should be kept in a fielded form, such as a spreadsheet, CSV file, or tab-delimited file. Auxiliary information necessary to interpret the metadata - such as explanations of codes, abbreviations, or algorithms used - should be included as accompanying documentation. To increase the findability, the Partners will also include keywords or key-phrases describing the subject or content of the data including relevant terms of the field.

As for the documents produced within the Project, including reports, following the Consortium Agreement and the D.1.1 Project Management Handbook, as well as the procedures agreed in D1.2 – Quality Assurance and Risk Management plan:

- A structured repository of Project documents, including restricted information has been developed. For Project-internal data sharing, such as the sharing of working documents, reports and deliverables, the Project uses SharePoint with restricted access to efficiently manage the Project information amongst the Project Partners and to enable the preservation of Project data and appropriate versioning of the documents.
- All Project documentation needs to conform to specific templates.
- Recommended document naming convention has been developed. The naming convention for all documents to be produced within the Project is provided in Section 4.3.1 of the D.1.1 Project Management Handbook.
- It is prescribed to use versioning property when modifying a document uploaded in the Project document repository or when producing different versions of code.

- Every document circulated to other Partners in the consortium should include a version number and date.
- When multiple contributors need to work on a document, it is recommended to use online documents that allow synchronous co-editing.

The research data which will be published should contain include the reference period, Project funding information (e.g., EU logo and information about the Grant Agreement and the action/program that funds the Project, official Project name and Project ID), release policy including dissemination rules, information about the collection of the data such as the data source, geographic coverage of the data, language, and file format.

4.2 Making data openly accessible

1) Medical images and health data

The consortium will develop the Federated Repository as a federated data storage solution that allows the collection and the exploitation of the INCISIVE autonomous/decentralized data sources (including mainly, but not only, medical images) in a transparent way. This will be achieved through the implementation of a data abstraction layer and uniform user interfaces enabling the user to perform queries over the INCISIVE data sources, regardless of their heterogeneity. In parallel, the consortium will implement a data donorship schema that will enable users to contribute with their data to the Federated Repository. The mechanism will allow the data providers to easily connect their data sources to the INCISIVE platform, assign the level of accessibility and contribution. It will also allow a user friendly/transparent opting-out procedure confirming with established directives. The details of this solution will be produced within WP5 of the Project.

Thus, the Federated Repository is fundamentally a data sharing platform with public access as one of its foundations. The development of this central milestone of the Project, drives other tasks, to ensure an “end product” of the Project - a secure and transparent data sharing mechanism available via the Federated Repository. The consortium will produce a detailed deployment strategy and operational roadmap for the Federated Repository (T8.4).

2) INCISIVE needs assessment study data (WP2)

The results of the WP2 surveys will be a part of a report which will be made available to the INCISIVE research team at KU and the Consortium Partners. KU will also be disseminating the findings of the study via journal articles and at relevant research conferences. A summary of the results will be available to any participants who request it. It will not be possible to identify participants from any such publications, as results will be anonymous (unidentifiable) and aggregated for the whole participants' group.

3) Other

Other data and publications generated under the Project will be disseminated in accordance with the Consortium Agreement. The Project publications and other dissemination and communication material, as well as the public Project deliverables will be made available via the Project web site and related platforms such as Zenodo. In public deliverables all personal data will be anonymized. Furthermore, the Partners will ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results. Within the limits of privacy laws and intellectual property protection, the digital research data generated in the action will be deposited in Zenodo or Open Aire in accordance with the Horizon 2020 Open Access policy.

These limits include restriction on disclosure different types of data that can either be used to identify individuals or that are of a commercially sensitive nature. Consequently, personal data of Project Partners or other stakeholders, raw qualitative research data from interviews, focus groups and workshops, draft reports, unfinished work, personal notes, plans for future research, preliminary analyses, peer reviews, and communication outside of a test setting, fall outside of the scope of the open access strategy.

4.3 Making data interoperable

1) Medical images and health data

For retrospective data, the Partners will focus on harmonising the input data from the different Data Providers and in this way make it available and integrated, as well as interoperable for the purposes of subsequent research challenges within the project, i.e. allowing re-use of this data by the researchers within the consortium, although they are datasets coming from different origins.

As for the Federated Repository, it will respect compliance with well-established standards, identified in Section 3.7.2 above, enhancing its interoperability aspects, and enabling its communication with existing systems. Under specific task within the Project, the responsible Partners will aim to will transform the data in an interoperable format. They will specifically undertake the alignment of the data in the Federated Repository with corresponding standards that will guide the whole development procedure. In specific, the standard on the risk management in medical devices (ISO 14971) as well as IEC 62304 (Medical device software — Software life cycle processes) will be considered, while HL7 FHIR and DICOM will be followed for interoperability and data exchange issues between INCISIVE and external systems. Finally, SNOMED, LOINC or other suitable semantic terminologies are under investigation to be used for data formalization. This will further support interoperability of the data in the Federated Repository with other efforts in this area.

1) INCISIVE needs assessment study data (WP2)

Not applicable.

2) Other

As for other data, the (meta)data that will be made open and re-usable will be in line with most widely used terminologies, standards, and methodologies to facilitate interoperability.

From a practical perspective, standard file formats will be used, considering the following guidelines:

- Non-proprietary and not tied to specific software,
- Open, documented standard,
- Common format used by the scientific community,
- Standard representation (Unicode, ASCII),
- Unencrypted,
- Uncompressed (where possible).

4.4 Increase data re-use (through clarifying licences)

At this stage of the Project, it is too early to provide details about licensing and re-use of data, thus in this version of the DMP we will focus on the intentions of the Partners and principles which will be followed.

1) Medical images and health data

The definition of the data donorship schema and the related IPR and licenses are part of the research work that will be done by the consortium. In particular, possible exploitation routes and general terms of use of the data stored in the Federated Repository will be a result of the works under Innovation management, business, exploitation and sustainability planning work package (WP8). They in turn be aligned with the Data Donorship Legal Framework (WP7, T7.4) which will form legal guidelines and terms, enabling external parties to safely donate their data in the INCISIVE.

Taking into the account the medical nature of the data which will be stored in the Federated Repository, the developed rules of re-use of data will take into consideration the legal guidelines arising from, inter alia, recommendations of the European Data Protection Supervisor (“EDPS”) Preliminary Opinion 8/2020 on the European Health Data Space. Many of the elements of those recommendations are relevant to the Federated Repository.

Furthermore, also from the IP ownership perspective, details on licensing of the data in the Federated Repository will be provided in the abovementioned data donorship mechanism.

2) INCISIVE needs assessment study data (WP2)

N/A

3) Other

At this early stage of the Project the following data that can be reused, at least, based on the current (initial) Project knowledge:

- System telemetry and performance data from INCISIVE applications will be published as Open-Data, to be retrieved for comparison and benchmarking for the High-Performance Computing community. Consortium Partners will set-up a data portal or repository (e.g., GitHub) for community to retrieve the datasets.
- Patient health and image data from the pilot studies are planned to be shared after the Project, anonymized/pseudonymized in the Federated Repository. Details will be determined at a later stage.
- All public deliverables and other public material (publications, communication material, etc.) will be accessible through the Project web site.

All the open research data will be made available for re-use without any data embargo, meaning that all data will be made openly available and free to re-use upon their publication. There will be no restriction on the use of data by third parties after the end of the Project. The data will remain reusable (labelled accordingly with the applicable licenses e.g., Creative Commons) forever.

5 Allocation of resources

5.1 Roles in data management

The main coordinating roles in data management are provided for in the Consortium Agreement and Grant Agreement as follows:

- Ethics and Legal Manager (TLX): is responsible to ensure that an appropriate data management plan is developed and used to protect the privacy of data and address all other data management aspects.
- The Innovation and Exploitation Manager (WR): is responsible to manage the knowledge produced during the Project lifecycle; manages execution of the overall exploitation plan of the Project and supports the Partners in setting up their individual business plans, in order to exploit the Project results.
- The Communication and Dissemination Manager (FTSS): is responsible to raise public awareness and ensure wide communication of the Project results and will also be responsible for the coordination of the scientific dissemination, clustering and standardization activities.

Furthermore, the WP/Task leaders will be expected to provide first level of data management within the scope of their role and ensure that the data of their WP/Task are treated according to the agreed Project principles and processes.

5.2 Resources

The costs related to making the data of the Federated Repository FAIR have already been budgeted in the INCISIVE consortium budget, e.g., costs of work for making the data interoperable, harmonization of data, etc. These costs are included in the overall budget of the respective Project Partners. The estimation of costs related to the sustainability of the Federated Repository in the post-Project period and investigation of possible ways to cover for these costs (business models) will be part of the WP8 activities and will be discussed in future updates of this DMP.

As described in the Description of Action, the Project results will be published mainly at fee-based open access scientific journals, following the OA Gold method, due to the high impact associated with certain journals. There exist many open access high-impact journals in the disciplines of optical networks and communications, published by IEEE, OSA and Elsevier allowing a variety of publication venues. For this reason, costs for publication fees have been foreseen in the consortium budget.

It is further anticipated that INCISIVE researchers will occasionally also follow the OA Green method in the case of conferences and workshop contributions, since the two OA methods are non-mutually exclusive. In that case the published article or the final peer-reviewed manuscript is archived by the researcher in an online scientific repository before, after or alongside its

publication. The authors must ensure open access to the publication within a maximum of six months. The Open Access Infrastructure for Research in Europe will be explored to determine which repository to choose (http://www.openaire.eu). Moreover, INCISIVE will exploit other support infrastructures provided by the EC towards data preservation, e.g. the Horizon Results Platform.

As the Project progresses and the different types of results are produced, it will be possible to provide further details on approaches for long-term preservation and accessibility of each type of dataset beyond the end of the funding period.

6 Data security

1) Medical images and health data

Initially, the retrospective data required for the development and training of the AI Tool will be stored in an infrastructure provided by FTSS IT facilities in Barcelona, Catalonia, Spain, EU ("**Temporary Infrastructure**"). As this activity of temporary storage was not initially foreseen in the DoA, it may require an amendment to the Grant Agreement and, in any case, at the stage of compiling this document, it is subject to the approval of the EC. The technical Project coordinator with the assistance of security partners have reviewed the description of the security features of the solution to ensure that storage and access conditions comply with the security standards required for the protection of personal information. It should be noted that FTSS is already providing data storage and data management services to the Catalonian Health System.

The Temporary Infrastructure solution is secured across several dimensions:

- The FTSS IT facilities is a huge secured system. Within the corporate network, a set of zones with different levels of security are defined. Access allowed or denied between the equipment in each of these areas is regulated by various firewalls or firewalls that limit the perimeter of the areas. The global infrastructure has four levels of security plus the platform security.
- Perimeter security guarding access to the cluster itself managed by FTSS and based on Kerberos. Kerberos is a computer-network authentication protocol that works based on tickets to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner.
- Data protection in the cluster from unauthorized visibility, sharing and manipulation based on Apache Ranger.
- Defining what users and applications can do with data based on Apache Atlas.
- Reporting on where data came from and how it is used based on Apache Atlas.

With level 3 security, Cloudera cluster is ready for full compliance with various industry and regulatory mandates and is ready for audit when necessary. The secure enterprise data hub (EDH) is one in which all data, both data-at-rest and data-in-transit, is encrypted and the key management system is fault-tolerant. Auditing mechanisms comply with industry, government, and regulatory standards, and extend from the EDH to the other systems that integrate with it. Cluster administrators are well trained, an expert certified security procedures and the cluster can pass technical review.

The infrastructure is a completely isolated system and external access is performed via VPN. The connectivity architecture is based on Cisco IPsec technology that enables the establishment of secure communications between users and the platform.

FTSS will securely store and make this data available to Consortium Partners for their research work during the first months of the Project. This solution will be deployed until the final federated solution for the storage of medical images and data is developed.

As regards this final solution, the images and accompanying data will be stored in Federated Repository developed within WP5. The details of this solution as well as scope of data stored therein will be produced as part of the future Project work and will be presented in next updates. Thus, at this point we present the intended solutions regarding storage and security features of this repository.

The Federated Repository will be developed using state-of-the-art technologies for secure storage, delivery and access of personal data, as well as managing the rights of the users. The details of this solution will be produced as part of the future Project work and will be presented in next updates.

In particular, the most suitable encryption method will be selected, according to the needs, and the Federated Repository's architecture, while also taking into account each individual hosting Partner's infrastructure. During data access procedures the data may be transmitted over an untrusted network. In order to make this data transfer secure, an encrypted connection e.g. in the form of an encrypted Virtual Private Network (VPN) shall be utilized. Should this VPN network be required to access the stored data, it will also provide a first level of access control. However, considering the needs of INCISIVE, a more sophisticated access control method is required, to set specific rules considering each user's rights and permission management. Access control entails authentication and authorization. An identity provisioning service shall be utilized, in order to authenticate users from various organizations and also set their access rights.

Each access to sensitive data must be logged in a trusted way, so no one can alter access logs. A blockchain based auditing mechanism will be adopted, with built-in smart contract capabilities, for logging all the critical actions done on the data stored on the cloud. This involves creating a log block after each data access, containing relative file and operation metadata, which is then broadcasted among all involved peers. Received blocks are validated and added to the blockchain, using adjacent blocks' hashes. This constitutes a secure and trusted notification and logging system by ensuring the tamper-proof, decentralised logging, auditing, and tracking of such actions leading to traceability and non-repudiation. Moreover, the blockchain mechanism will allow for logging, auditing and tracking of the transactional data activities, e.g., data submission, sharing, access, preserving at the same time confidentiality and data access permissions.

Furthermore, having investigated involved nodes' capabilities, specialized hardware security features will be utilized. A prominent example is Intel's Software Guard Extensions (SGX), which aims to define protected areas of the memory, called enclaves. In those enclaves the data is encrypted and can only be accessed by certain processes. As a result, processes outside the enclave have no access neither to the data, nor to the instructions executed. This rules out the possibility of data leaks due to attacks within the operating system of participating nodes.

In this way, there is a sufficient level of assurance that the accessed, delivered, stored and transmitted content will be managed by authorised persons, with well-defined rights, at the right time.

2) INCISIVE needs assessment study data (WP2)

All data will be entered, stored and backed-up in a secure manner by the research associate for the INCISIVE Project at Kingston University. Once the study has been finalised, all personal/identifiable information will be removed as per KU data protection policy. Finalised and fully anonymised study data upon completion of the study will be stored for 10 years as per Kingston University research data management policy and retention policy. When reporting the results of the study, no information will be released which will enable the reader to identify who the respondent was.

3) Other

The following mail research datasets identified at this point will be stored locally, by each relevant Consortium Partner according to their internal rules. Indicative examples follow:

- UX design workshop data and results of evaluation of INCISIVE - finalized and fully anonymized (unidentifiable) data upon completion of the task will be stored for 10 years as per Kingston University research data management policy and retention policy;
- System telemetry and performance data from INCISIVE applications data will be stored in BSC clusters and supercomputers. This IT infrastructure has its own security and data plans to avoid intrusions and breaches. Also, access is firewalled, and private credentials are required. The system is by itself not accessible from external networks;
- AI algorithms and analysis of their performance data will be stored both in the Temporary Infrastructure and locally by the technical Partner developing them, secured in compliance with previous task decisions and data protection regulations;
- Interviews and/or surveys will aim to validate and improve the INCISIVE business models will be stored in Project's Dropbox folder accessible by WR's employees and in the Project's respective SharePoint folder. All employees of WR have signed relevant NDAs. Data will be kept for the whole duration of the Project.

When processing any research data on their local infrastructure, the Partners will observe general data protection principles regarding data security. In particular, each Partner undertakes to:

- Follow the agreed pseudonymization and anonymization guidelines,
- Keep pseudonymized data and pseudonyms of respondents separate;
- Use their available local file servers to periodically create backups of the relevant materials.
- Encrypt data if it is deemed necessary by the local researchers;

- Store data in at least two separate locations to avoid loss of data;
- Limit the use of USB flash drives, with a clear commitment not to store any personal data on such sticks;
- Save digital files in one the preferred formats (see attached table), and
- Label files in a systematically structured way in order to ensure the coherence of the final dataset.

Additionally, all other relevant documentation created during the Project such as deliverables will be self - archived and preserved in INCISIVE SharePoint that has been created for the purposes of the Project. It allows users to store files in the cloud, share files, and edit documents, spreadsheets, and presentations with collaborators. The INCISIVE SharePoint is accessible to all of the Partners of the INCISIVE consortium.

7 Ethical aspects and intellectual property rights

7.1 Ethical Issues

The INCISIVE Partners have committed to comply with the ethical principles as set out in Article 34 of the Grant Agreement, which, among other, states that all activities must be carried out in compliance with:

- Ethical principles (including the highest standards of research integrity)
- Applicable international, EU and national law.

The ethical aspects of the Project will be assessed under WP7, which sets out the ethics requirements that the Project must comply with. More specifically under T7.2 the required ethical approval actions will be prepared and performed by clinical partners with the help of the appropriate authorities and other Partners when necessary. The composition of the formal Ethics letter, as well as the process for obtaining the formal approvals will take place within this task.

Additionally, the Project partners confirm to respect the EU and national law requirements on privacy and data protection and to adhere to the research ethics standards applicable to Horizon 2020 research. In accordance with the data minimization, data retention and purpose limitation principle, personal data will not be collected beyond the scope of the processing objectives and will not be stored for longer than necessary.

In the context of the Project, the ethical aspects relate primarily to the collection and use of medical images and health data and interviews with patients and HCP.

1) Medical images and health data

Each of the Data Providers will obtain required ethical approvals for the extraction of retrospective data and its use in the context of the Project. At the time of preparation of this document, five Data Providers obtained their ethical approvals for WP3 (retrospective data) including: AUTH, DISBA, UNITOV, UNS and GOC. The following Data Providers are still in the process of obtaining approvals:

- HCS obtained ethical approval from their institution and one other hospital; they are still waiting for the approval from the second hospital which is estimated by the end of March 2021.
- UoA has submitted the proposal to the first ethics committee to which approval has been granted; the final approval is estimated by the end of March 2021.
- IDIBAPS has been submitted the proposal to the ethics committee with initial feedback suggesting additional clarifications about issues related to anonymization and data storage. IDIBAPS is currently contacting AUTH and FTSS about these issues in order to resume with the application process.

Moreover, the data protection officer ("DPO") or – in the absence of a formal DPO appointment at the relevant site – a designated privacy person for each Data Provider was contacted to check if the provision of the retrospective data for the purposes of the Project complies with the local rules and whether any additional requirements for its processing are necessary from the perspective of personal data protection laws.

From a GDPR compliance perspective, the Consortium Partners involved in the collection and use of the retrospective data will be bound by data sharing agreement which will set out their respective tasks and obligations as considered joint controllers of this data.

2) INCISIVE needs assessment study data (WP2)

Ethical approvals for all studies required to complete T2.1 were obtained by the task leader KU. Data Providers requiring extra layer of ethics at their corresponding institutions: DISBA, UNITOV, GOC, UNS and IDIBAPS got their ethical approvals for the different studies involved in T2.1. Data providers such as AUTH, HCS and UoA did not require any extra layer of ethics, hence they were covered by KU ethics.

3) Other

At later stage of the Project, the relevant Data Providers will obtain ethical approvals for pilot studies in each country within T6.4.

7.2 Confidentiality

All INCISIVE Partners must keep any data, documents or other material confidential during the implementation for the Project and for four years after end of the Project in accordance with Article 36 of the Grant Agreement. Further detail on confidentiality can be found in Article 36 of the Grant Agreement.

7.3 IPR

Issues regarding the protection of intellectual property rights (IPRs) and confidential information were addressed in detail within the Consortium Agreement. In particular, the Consortium Agreement regulates the IP-Ownership, Access Rights to Background and Foreground IP (Articles 8, 9 and 10). Moreover, in accordance with Article 24 of the Grand Agreement, Background was identified for all Partners, if applicable. In the first months of the Project, IPR control spreadsheets have been circulated and existing IP (Background), foreground IP and contributed assets were identified by the Partners. The details will be described in the innovation strategies (D8.1, 8.3 and 8.6).

8 Conclusions

The document presented initial INCISIVE Data Management Plan based on the initial data sets identified by all Partners. The DMP will be revised and updated during the entire duration of the Project. The DMP will be updated at least by the mid-term and final review to fine-tune it to the data generated and the uses identified by the consortium since not all data or potential uses are clear from the start. New versions of the DMP will be created whenever important changes to the Project occur due to inclusion of new data sets, changes in consortium policies or external factors.

9 References

1. European Commission, Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 26 July 2016.
2. European Commission, Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020, Version 3.2, 21 March 2017.
3. European Data Protection Supervisor Preliminary Opinion 8/2020 on the European Health Data Space.

10 Annex

1. Responses to DMP questionnaire (as of 12 March 2021).