INCISIVE

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 7.5

Final Data Management Report

WP7 – Legal and Ethics Management

29-03-2024

Revision 1.0

Status: Final

Grant Agreement n 952179

European Commission

DOCUMENT CONTROL	
Project reference	Grant Agreement number: 952179
Document name	D7.5 Final Data Management Report
Work Package	WP7
Work Package Title	Legal and Ethics Management
Dissemination level	PU
Revision	1.0
Status	Final
Reviewers	Lithin Zacharias, Shereen Nabhani (KU), Tatjana Loncar-Turukalo (UNS)
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	All Partners

REVISION HISTORY				
Revision	**Date**	**Author**	**Organisation**	**Description**
0.1	22/9/2023	Magdalena Kogut - Czarkowska	TLX	Structure and scoping
0.2	20/2/2024	Magdalena Kogut - Czarkowska	TLX	Outline and first draft for input by the partners
0.3	14/3/2024	Magdalena Kogut - Czarkowska	TLX	Version for peer-review
0.9	26/3/2024	Magdalena Kogut - Czarkowska	TLX	Pre-final version after peer-review

REVISION HISTORY				
Revision	**Date**	**Author**	**Organisation**	**Description**
1.0	29/3/2024	Magdalena Kogut-Czarkowska	TLX	Final version 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This deliverable contains original unpublished work except where clearly indicated otherwise.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

Table of Tables

Table of Figures

Terms and Abbreviations

Term	Description
AI Toolbox	AI solutions to improve cancer detection systems and enhance the clinical workflow
Anonymisation	Anonymous information is one which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly (recital 26 GDPR). Anonymization means that an individual (person) cannot be re-identified with reasonable effort based on the data provided or by combining the data with additional data points.
Central infrastructure	The cloud infrastructure of INCISIVE, hosted by Azure, comprised by 4 Virtual Machines, located in Central France that contains the centralized services required to make the INCISIVE Platform work.
Central Node	Central data storage space that hosts Data from the Data Providers, technically acting in the same way as the Federated nodes. The Central Node can be used as one node of the several INCISIVE Federated Nodes and/or as a centralised data repository where AI training or federated learning can take place.
Data Access Committee or DAC	The committee to review and make decisions as to whether to approve a potential application for re-use of Data of the INCISIVE Project by external Data Users.
Data User	The entity or a person who uses the Data available in the INCISIVE Repository.
EUCAIM or EUCAIM Project	EUropean Federation for CAncer IMages project co-funded by the European Union under Grant Agreement 101100633. More details: https://cancerimage.eu/
Federated Data Sharing	The Data Providers share the Data by keeping them within their infrastructures (physical or virtual). After being pre-processed locally, they are made interoperable with other data existing in the INCISIVE Repository, through a data sharing mechanism.
Federated Learning	Means that AI model is trained in a distributed way using the required Federated nodes, i.e., the nodes that have the Data that matched with the user query. Each Federated or Central

	Node contains a particular set of Data that may be required for training, and it does not leave the node to ensure privacy. The model is trained in each Federated node, including Central Node, and then it is sent to the Central infrastructure to be merged to gather ‘central knowledge’. This training-merging process can be repeated more than once for the same model as the more times this is done, the more robust is the solution.
Federated Node	Dedicated infrastructure (physical or virtual) in which the INCISIVE Data Provider stores the Data which they contribute to the INCISIVE Repository.
Federated Space	A virtual space formed by the composition of all Federated nodes, including the Central node, enabling the performance of centralized operations, such as the training of the AI models, access to Data utilized in the INCISIVE Platform, and visualizations of the results etc.
Hybrid data sharing	Data sharing which takes place using both Federated data sharing and a Central node.
INCISIVE Data or Data	Anonymized medical data and images made available for re-use in the INCISIVE Repository, initially collected by INCISIVE Data Providers for the purposes of INCISIVE Project.
INCISIVE Data Providers or Data Providers	The following INCISIVE Partners: AUTH, HCS, UoA, UNITOV, DISBA, GOC, VIS, OIV, IDIBAPS.
INCISIVE Partners	Beneficiaries of INCISIVE Project
INCISIVE Platform or Platform	Platform (technical infrastructure integrating several components) provided by the Project which includes the Hybrid repository, AI development workspaces for AI training and the Inference services.
INCISIVE Project or Project	The INCISIVE project, Grant Agreement n 952179.
INCISIVE Repository	collection of INCISIVE Data and technical infrastructure (subset of the INCISIVE Platform) for the data’s secure storage with the aim of making the data available for re-use by the Data Users under the conditions defined in this document
Inference services	The process of using the AI models over new input data to obtain the targeted results, e.g., predictions or tumour segmentations.
Permitted Research Purposes	Purposes for which the INCISIVE Data may be used by Data Users

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)
Temporary Infrastructure	IT infrastructure stack hosted by FTSS
Abbreviation	**Description**
EC	European Commission
TCC	Technical and Clinical Committee
WP	Work Package
DMP	Data Management Plan
HCP	Heath Care Professional
GDPR	General Data Protection Regulation (regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. The wordings ‘personal data’, ‘data subject’, ‘processing’, in this document have the meanings set out in the GDPR.
MAG, ICCS, CERTH, CeRICT, BSC, ED, TIS, SQD, AUTH, UNS, VIS, DISBA, FTSS, HCS, PASYKAF, UNITOV, UOA, UH, IDIBAPS, GOC, KU, CUT, TLX, WR, MDT, ADAPTIT	INCISIVE partners’ abbreviations as defined in the Grant Agreement number 952179 – INCISIVE OIV is Oncology Institute of Vojvodina, established in Put dr Goldmana 4, 21204 Sremska Kamenica, Serbia, linked to UNS (a project partner) as a research base of the Medical Faculty of UNS

1 Executive Summary

Data Management Plans (DMPs) are 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.

Thus, the main goals of the Final INCISIVE Data Management Report (Final DMP) are to outline:

- the datasets collected/generated, including the context and procedures of the collection/generation and ethical and legal requirements.
- the measures and procedures for FAIR (findable, accessible, interoperable, reusable) data.
- the measures for the adequate management of the data from an ethical and a security point of view.

INCISIVE project collected and generated various data to meet its objectives to:

- examine user needs and requirements with regard to the INCISIVE Platform;
- collect de-identified and verified medical data for the creation of INCISIVE Repository;
- develop AI tools for cancer detection and
- validate those tools in feasibility studies.

The scope of this Final DMP is to describe the data management life cycle for all datasets collected, processed and/ or generated in all Work Packages (WP) within the INCISIVE project, with particular emphasis of anonymized medical images and health data (Pan-European Repository of Health Images) which one of the core results of the project. The methodology proposed by the European Commission Guidelines has been adopted for the deliverable compilation.

The first version of INCISIVE DMP (D7.1) was submitted in M6. An update to this document had been delivered in September 2021. This second version of D7.1 provided a more detailed data management plan, explaining the use of retrospective data in the project. Next, two more iterations of the plan were prepared in each year of the project (March 2022 and April 2023). This deliverable builds on the work performed during project term and comprises the final version of the DMP. The most relevant updates in this version, when compared with Initial Data Management Plan (D7.1), include:

- Details related to collection, FAIR-ification and management of medical data required to populate the INCISIVE Repository, including anonymization and preparation of this data to be shared and re-used beyond the duration of the Project;
- Detailed description of other datasets that have been collected, used or generated by the project from M1 to M42 in each task (Annex 1);

- Explanation about the standards and utility assurance processes implemented by the Project;
- Description of approach to making the data collected during the Project FAIR;
- Planned cooperation with EUCAIM in order to support making INCISIVE medical data re-usable for external researchers and available to broader research community;
- Alignment with other deliverables and results of INCISIVE project.

2 Introduction

2.1 Purpose and scope

Purpose of this final 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 has been collected, processed or generated by the Project,
- Provide an analysis of the main elements of the data management policy, which were used by the Partners regarding all the datasets which were generated by the Project,
- Provide details about and guarantee the preservation of the data collected during the Project, as well as any results derived from the associated research,
- Inform how the INCISIVE consortium addressed the ethical issues related to data, which were collected during the Project timeframe.

The INCISIVE DMP evolved during the lifetime of the Project, with its initial version submitted as D7.1 and then updated several times during the course of the Project. More specifically, updates of the DMP were presented in March 2022 and April 2023. This document summarizes the work on data management performed through the entire Project.

2.2 Document structure

The rest of the document is structured as follows:

- Section 3 provides a brief description of data sets which were 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 alignment with the sustainability of the project outcomes.

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
- D7.3: Data Donation Legal Framework: defines the rules and standards regarding the data donation in the healthcare sector and form guidelines and terms of such donorship required to put forth the rules of sharing of data in the INCISIVE Repository.
- D7.4: IPR Management Report: IPR management report, which is 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'.[1] 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 results of experiments, measurements, observations resulting from fieldwork, statistics, 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 was collected and processed during the Project for the purposes of development of the main results achieved from INCISIVE's research:

- The INCISIVE AI-driven models enhancing image processing and data analysis focusing on improving sensitivity and specificity in diagnosis and statistical assessment of cancer, as described in D4.3
- The INCISIVE Pan-European Repository of Health Images that enables the secure access and sharing of data and ultimately allow the large-scale adoption of such solutions in cancer diagnosis and follow-up as described in D5.3

[1] 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.

- The INCISIVE platform which through its HPC and HDPA-as-a-service, provides secure and cost-effective performance of computationally intensive processing, without the need for maintaining expensive equipment, as described in D4.3
- 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, as described in D4.3
- The INCISIVE anonymization mechanism aiming to enable legal/ethical sharing and processing of medical data, as described in D5.2 and protocols developed for post-project data anonymization
- A pool of scientific publications in high ranked conferences and high impact journals. List of the publications is provided in D9.6

3.3 Methodology of work

The specific data sets for the Project needed 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 would 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?

In the last year of the project, the Partners were additionally asked the following questions:

6. FAIR DATA

a) To whom might the data be useful (data utility)?

b) Data supporting publications and required tools. Does the described data support any scientific publication? If yes, please indicate which. Data supports publication: YES/NO. Is there a data availability statement provided along with the publication?

c) Making data findable, including provisions for metadata. Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)? Is data discoverable/Persistent identifier (PID)/Search keywords/Versioning/Metadata creation;

d) Making data openly accessible. [Indicate details asked below] Data openly accessible: YES/YES, but only to project partners/NO. How it will be accessible/Methods/software tools to access data/Repository/Restrictions on access;

e) Making data interoperable. How is the data made interoperable? Indicate details: Interoperability/Data and metadata vocabularies/Use of standard vocabularies/Mappings to commonly used vocabularies;

f) Increase data re-use. Licence/Availability for re-use/Usable by third parties after end of project/Re-use timeframe/Data quality assurance process;

g) Allocation of resources. Are the costs covered by INCISIVE project?

Moreover, responses were provided in the first months of the Project and then gradually updated, as the Project progressed. Total of 4 iterations of the DMP questionnaire were prepared. The attached Annex 1 contains the final responses provided by the partners, grouped by specific data sets processed under the tasks they are involved.

3.4 Data sets and data format.

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

1) Medical images and health data

During the course of the project, medical images and health data (encompassing retrospective training data, prospective training data, prospective data collected in observational and feasibility INCISIVE pilot studies and open-source data) were used to accomplish various goals of the Project, in particular to train the AI Toolbox, to develop final INCISIVE Repository and develop the donorship mechanism.

a. Retrospective and prospective training data

The following Partners: AUTH, HCS, UoA, UNITOV, DISBA, GOC, UNS, VIS, IDIBAPS (‘Data Providers’) provided de-identified health and medical images data which were used for purposes of subsequent research tasks within WP3, WP4 and WP5. These consisted of retrospective training data and prospective training data.

In first stage of the project, retrospective training data were extracted by each Data Provider from their hospital systems and uploaded to Temporary Infrastructure (as further described in Section 6 below). This data collection started in May 2021. The data comprised a wide variety of patient data and existing cancer images (CT, MRI, PET/CT, Ultrasounds, X-Rays, Mammographs etc). Images were provided in DICOM format.

In addition to the medical images, data, additional patient information was collected in .xls 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.) was collected for all cancer types at distinct time points (the timepoints considered are: baseline, after 1st treatment, 1st follow up, 2nd follow up).

Consortium specified the formats and categories of data and defined them in Data de-identification protocol and guidelines.

The retrospective training data enabled the excessive training of the foreseen AI Toolbox. The exact scope of personal data processed was defined in a protocol. Further information about INCISIVE approach to collecting and processing retrospective training data (also referred to as 'retrospective data') is provided in Section 7.2.

The retrospective data was combined with prospective training data. Prospective training data involved additional collection of data from the patients receiving diagnoses and/or treatment from the Data Providers. Each of the Data Providers identified such patients and obtained their consent for including their data in the study. The protocol for this data collection was discussed and decided by all Data Providers. Respective partners have applied for and obtained approvals from the ethics committees. The detailed information about the obtained approvals was reported in D7.2, which also contained the informed consent forms and information sheets for the prospective studies.

The data collected in terms of patient numbers for retrospective and prospective for AI training studies were the following:

Cancer Type	Number of patients-retrospective	Number of patients-prospective
Lung	2962	65
Breast	4701	77
Colorectal	890	39
Prostate	217	157

Table 1: Patient numbers for retrospective and prospective data

INCISIVE pilot studies' (observational and feasibility) data collection was aligned with previous data collection regarding both formats and categories of the data. Partners acting as DPs have obtained ethical approvals for data collection and data sharing, which included both institutional approval and patient consent for in project and post-project usage of de-identified data. Data collected in observational study served for the quantitative evaluation of the AI toolbox, in terms of AI model performance metrics (depending on the model type, e.g. accuracy, sensitivity, specificity, F1, etc).

The project collected the following numbers[2] of prospective observational data on Breast, Lung, Colorectal and Prostate cancer cases:

OBSERVATIONAL STUDY

BREAST CANCER OBSERVATIONAL DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	200	1066	1328
SERBIA	50	35	107
ITALY	450	296	440
CYPRUS	155	32	139

LUNG CANCER OBSERVATIONAL DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	231	38	116
ITALY	350	269	357
CYPRUS	100	14	37

COLORECTAL CANCER OBSERVATIONAL DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	231	24	76
ITALY	350	250	360
CYPRUS	50	20	48

PROSTATE CANCER OBSERVATIONAL DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases

[2] Valid as of 26/03/2024.

GREECE	50	50	50
SPAIN	107	117	117
CYPRUS	155	70	183

Table 2: Detailed cases and patient numbers for observational data.

Final numbers will be provided in the final technical report.

The feasibility study included data collection to evaluate the INCISIVE platform and its AI services, with the aim of assessing acceptability, usability, trust, and satisfaction The data collection has started in July 2023. This data has been used for evaluation of AI services, including segmentation and localization tools and has not been annotated.

The project collected the following numbers of prospective feasibility data on Breast, Lung, Colorectal and Prostate cancer cases:

FEASIBILITY DATA COLLECTION[3]

BREAST CANCER FEASIBILITY DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	100	117	129
SERBIA	20	24	34
ITALY	100	74	109
CYPRUS	20	20	20

LUNG CANCER FEASIBILITY DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	75	20	72
ITALY	300	225	323
CYPRUS	20	10	20

COLORECTAL CANCER FEASIBILITY DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	

[3] Valid as of 26/03/2024.

	cases	patients	cases
GREECE	100	14	43
ITALY	300	237	309
CYPRUS	10	8	10

PROSTATE CANCER FEASIBILITY DATA COLLECTION

COUNTRY	GA & D2.6	COLLECTED	
	cases	patients	cases
GREECE	25	23	23
SPAIN	70	70	70
CYPRUS	25	14	25

Table 3: Detailed cases and patient numbers for feasibility study data.

Further update will be included in the final technical report.

As number of patients and number of imaging examination (cases) was dependant on the pace of patients visits and patient willingness to sign the consent, the overall numbers reached are higher than promised in the grant agreement, though in some cancer higher than expected while in some (such as colorectal/lung) for some DPs lower than expected.

All medical images and health data were de-identified prior to sharing and the algorithms did not take a patient’s name or other identifying metadata into consideration to analyse their medical images. Further explanation is provided in Section 7.2

b. Open data sources

The retrospective and prospective data sources were augmented with open-source data. Consortium has investigated 48 open data sources. The detailed list of the open data sources (including detailed description and legal restrictions for their use) was kept.

These open-source data were selected based on their relevance according to the Project objectives, their availability and compatibility with the retrospective data. They were digested and formatted based on the needs identified by the consortium. The proper acknowledgment of these open sources and of the work of others was guaranteed by applying appropriate citation and quotation methods. In particular, the consortium secured a permission to use the InBreast Database for the AI Toolbox training and explored several open databases that can be used without permission and has concluded to some that can be used to solve the clinical challenges that the medical professionals of the consortium have requested, such as the LIDC-IDRI, Lung PET-CT-Dx, and NSCLC Radiomics databases from TCIA.

In terms of implementation of classification models using radiomic analysis (as part of T4.6), AUTH used the following database NSCLC-Radiomics (available at:

https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics), which includes CT images with the segmentation of the tumour as well, as well as information regarding the stage of the cancer. According to the copyright note, the data can be used by the research community.

c. Patient data from validation sites (study pilots)

Patient data generated from the observational and interventional studies across the validation sites (T6.4/6.5). At each of the 8 pilot sites, the INCISIVE platform was used to store and evaluate the data from a number of patients. The data includes relevant health data such as clinical data, imaging data, histopathology data, aligned with INCISIVE pilots' protocol definition for each cancer type (T2.6).

2) INCISIVE needs assessment study data (WP2)

These data sets collected within WP2, 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 involved email semi-structured interviews with HCPs within the Project consortium. The interview schedule consisted of 28 questions exploring the care pathway in each country. After collection, data was 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 involved surveying HCPs using online data collection method. The survey consisted of 58 questions divided into 6 sections.
- For Patients: 'Perceptions and experiences of cancer patients regarding existing cancer care pathways across Europe'. The interview consisted of 15 open ended questions. Interviews were audio-recorded and then transcribed verbatim into text.

The HCP and Patient interviews were part of a report and deliverable D2.1 (submitted on 16 April 2021) which was be made available to the INCISIVE research team at KU and the project consortium partners.

3) Other

Apart from the principal categories mentioned above, the following main sets of data collected, processed and/or generated in the course of the Project:

a. WP2

- Outcomes produced during UX design workshops with participation of INCISIVE stakeholders (T2.2). Design thinking method and Delphi approach was used as methodological approaches to better understand and prioritise users' requirements. The outcomes included conducting two Delphi studies: one with the primary users of the proposed INCISIVE AI toolbox (healthcare professionals involved in cancer care) and the other with the primary users of the INCISIVE repository. Performing the Delphi study involved a series of questionnaires to be administered to participants (primary users of the AI toolbox and primary users of the repository) to achieve prioritisation and consensus of the features and implementation barriers for the INCISIVE system. Related deliverable (D2.2) was submitted on 30 September 2021.

b. WP3

- Algorithm and methodology for data harmonization (T3.1). The result is the development of a methodology and algorithms that receives the heterogenous data from different sources and integrates them in a specific homogeneous structure and produces error reports.
- Software and data related to the technical requirements of the technology providers, container configuration files to set up the software integration environment (T3.2).
- 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 was 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).

c. WP4

- 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 were applied on de-identified patient data with the goal of improving the diagnosis on cancer diseases. Their performance was be evaluated on each case of cancer. Estimation models along with the aforementioned algorithms were used to estimate the progress of the disease for each patient (various tasks in WP4).
- ML performance data - Performance evaluation metrics
- Metadata that needed to be maintained for each AI Engine and AI Model. The full list of metadata was made available on project SharePoint.

d. WP5

- Metadata about the records of actions of the users on the data in the repository (Task 5.5.) The data collected and generated were used for generating the logs which can be inspected through the auditing mechanism. The role of the Transaction Tracker and the audit mechanism was to provide transparency.

- The metadata for describing health data that INCISIVE project has developed together with cancer data.
- Contact data (incl. Names, address, email address, phone numbers) of future Data Providers and Data users collected through the Data Sharing Portal. No data has yet been collected as this process will start mainly after the end of the project. This data will be kept confidential and not be FAIR.

e. WP6

- Analytics on AI models' accuracy, sensitivity, specificity, efficiency and performance evaluation results (T6.4).
- Analytics resulting from the usability evaluation of the federated repository including its search engine and workspace both for data providers and data users (T6.2)
- Interviews with clinicians and data generated by qualitative assessment (T6.5)

f. WP8

- Additional open-source and proprietary market intelligence related data for performing a thorough AI for Health Imaging Market Analysis (T8.1).
- Interviews and surveys aimed to validate and improve the INCISIVE business models and service value propositions as well as to establish direct contacts and relationships that may result in a more pro-active adoption of the Project's main exploitable assets (i.e., the AI Toolbox and the Repository) by target groups. Subjects were stakeholders that have been identified as potential early adopters and customers of the INCISIVE services (i.e., doctors and hospital managers), experts in the AI for health industry and members of the Advisory Board.

g. WP9

- 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).
- Stakeholders contacts from attendees' registration at INCISIVE events

h. Cross WP

- Project deliverables and management reports (public or confidential) as well as publications for scientific dissemination and other more general communication material generated by the Project Partners.

The specific data formats are described in Annex 1 in relation to each data set.

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

The Project relied on existing data sets as well as produced new, original data.

1) Medical images and health data

The training data provided by the INCISIVE Data Providers was augmented with open-source data. The consortium has investigated 48 open data sources. The detailed list of the open data sources (including detailed description and legal restrictions for their use) was kept. The following resources were re-used for the purpose of the following tasks of the project:

- NSCLC Radiomics (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics) for lung cancer CT scan segmentation
- PICAI Grand Challenge for Prostate Cancer MRI Segmentation (https://pi-cai.grand-challenge.org/)
- NSCLC Radiogenomics (https://wiki.cancerimagingarchive.net/display/Public/NSCLC+Radiogenomics)
- NSCLC-Radiomics-Genomics (https://wiki.cancerimagingarchive.net/display/Public/NSCLC-Radiomics-Genomics)

The open-source data were used separately from INCISIVE data and have not been integrated with INCISIVE data repository. Open-source data was anonymized and did 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 was 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

Data sets described above in Section 3.4 consist of original unpublished data, except where clearly indicated otherwise. The original data was collected by the Project Partners. Naturally, the data collected or developed during earlier tasks had fuelled the progress of subsequent tasks.

3.6 Expected size of the data

1) Medical images and health data

The size of retrospective training data is approx. 4.182 TB of data provided by 9 Data Providers. After the initial development stage, the AI Toolbox was 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.

Last, but not least, the ultimate objective of INCISIVE was to create the INCISIVE Repository, which stores digital images of various cancer types.

Overall, the approximate size of the data in the INCISIVE Repository is ~5.1 TB.

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 encompass:

- For HCPs (first round of interviews): responses from a sample of 7-10 oncology specialized HCPs within the Project consortium.
- For HCPs (surveys): 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 was 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, did not to exceed a maximum of 250 MBs.

Apart from the 'raw' heath data, user interviews, and reports INCISIVE Project also processed 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 also generated additional data.

Further details on the size of the respective datasets are available in Annex 1.

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 delivered a standalone INCISIVE 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 INCISIVE repository was built upon a Hybrid storage approach and a set of standardized open APIs that 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 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.

The results of the Project research activities and INCISIVE Repository established as a result of the Project are primarily 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 INCISIVE Repository contributes 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 informed the design, development and implementation of the INCISIVE Platform and aimed to benefit all stakeholders mentioned above, as they were a reliable source of information, new insights and methods. In particular, they can 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 examined 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. The details of this exercise per dataset are provided in Annex 1. Some of the notable examples of re-usable data include:

- AI training materials, which, although specific to INCISIVE, 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.
- The project has developed metadata for describing health data and the list of metadata that need to be maintained for each AI engine and AI model. This metadata can be reused by other researchers in other projects.
- The project developed data donorship framework, along with policies and terms of use of health data for research purposes.

3.7.2 Standards

INCISIVE made use of existing standards where applicable and made pre-standardization work suggestions contributing to existing efforts especially focusing on related ontologies, vocabulary, image labelling, annotation and de-identification. The following standards were identified and referred to in INCISIVE:

- Medical images
    - Were standardized with DICOM (in particular DICOM PS3.15 2021a - Security and System Management Profiles) and are stored in PACS.
- Clinical data
    - Were semantically encoded with SNOMED CT and LOINC standards.

    - Were syntactically structured with HL7 FHIR messages for each cancer type. HL7 FHIR, in particular HL7 FHIR R4 – Fast Healthcare Interoperability Resource to exchange the data from different sources into a common data model.
- Clinical Report
    - Were standardized with HL7 CDA Standard for the output for each cancer type when running AI services for model inference.
- Other
    - IHE standard has been taken into account to define the INCISIVE use cases.
    - Risk management in medical devices (ISO 14971) as well as IEC 62304 (Medical device software — Software life cycle processes);
    - D3.4.Standardization Suggestions.
    - 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] were considered, as well as IEEE standards and ongoing standardisation activities[5]. INCISIVE also provided input to the FUTURE-AI framework for trustworthy AI (https://future-ai.eu/) which is the result of common work of all AI4HI projects. The framework promotes 6 principles and a number of recommendations for the training, validation and deployment of trustworthy AI (FUTURE stands for Fairness, Universality, Traceability, Usability, Robustness, Explainability). As next step, Consortium undertook to align its results 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 contributes to existing standards and stimulate creation of new ones.

More specifically, INCISIVE Platform enables a virtual environment for all Data Providers with all the necessary systems and databases to ensure the same standardized process from multiple sources. Clinical data and medical images are processed locally at the Hybrid Nodes (or Central Node, if chosen by the Data Provider) and must never leave their facilities, either to be processed or to be ingested into their local storage. The ETL tool running at each Data provider's node transforms the clinical data into HL7 FHIR messages and save the resources on the FHIR server,

[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

and the medical images that are uploaded with DICOM format are just saved on PACS. The use of HL7 FHIR messages and DICOM images ensures the harmonization of Data in a common data model to share them in a standard way and enable Hybrid AI training. In particular, the Data Providers fill clinical data into an Excel template, which is then used to feed the ETL tool. The same process takes place for medical images and annotations, which are compressed by Data Providers with DICOM files and NIFTI annotations and are also uploaded to the ETL tool of their Hybrid Node. It is the Data Provider's responsibility to use the quality check tool to validate that the uploaded data is correct and useful for AI training, and the de-identification tool for the medical images before the uploading process to the ETL tool.

The ETL tool transforms the Excel template into an HL7 FHIR message with SNOMED CT and LOINC codes to preserve the semantic meaning of all clinical information exchanged. It also processes and stores medical images and annotations from DICOM and NIFTI files. The message with all clinical data is stored on the local FHIR server and the medical images and notes are stored on the local PACS, both of which ensure the localization, accessibility, interoperability and reuse of the Data within each Data Provider's Hybrid Node.

AI researchers run the AI training models that return the result and display it in the UI, the data used by the AI training models is fetched and processed within each Node of the Hybrid infrastructure.

The Common Data Model (CDM), including the INCISIVE HL7 FHIR implementation, the clinical data semantic codification with SNOMED and LOINC, and the DICOM and NIFTI standardization process was explained in the D3.4.Standardization Suggestions.

Finally, the Project investigated legal ‘standards’ and precedents for data sharing in Task 7.4. The outcome of this work was reported in D7.3 Data Donation Legal Framework, which was submitted in M30.

3.7.3 Quality assurance

Quality Assurance took 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 were defined in D1.2 - Quality Assurance and Risk Management plan and are, therefore, will not be 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 was 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 have agreed on the data de-identification 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 has been ensured through a series of workshops organized between the technical Partners and the Data Providers. The workshops for all cancer types were concluded and a document named ‘Data Uploading Guidelines’ was drafted, containing all the data requirements for each cancer type and imaging modality to ensure the homogeneity of data generated through INCISIVE.

On top of that a Data Integration Quality Check tool was implemented under Task3.1 that checks the data for possible errors and informs the user on corrective actions prior to data upload, ensuring high data quality.

Within WP5 the relevant Partners contributed to the quality assurance of the INCISIVE data in the INCISIVE Repository through the implementation of data annotation, de-identification and data curation tools. The de-identification protocol used was based on the DICOM PS3.15 protocol by NEMA. The de-identification tool that was developed offers the ability to produce de-identified images with a varying level of de-identification.

2) INCISIVE needs assessment study data (WP2)

The quality of the WP2 data that was 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 were followed. These quality assurance procedures were further detailed and specified within each work package as the Project progressed.

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 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 main re-usable data produced by the project. We also provide a description on how those principles are enshrined in the INCISIVE Repository.

1) Medical images and health data

The retrospective data which was shared by the Data Providers came with their own metadata stemming from each of those providers. The Partners had to 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 was reached initially for retrospective training data, during the definition of the retrospective study protocol (see point 7.2 below). The metadata for the annotation of the prospective data was defined from the beginning, making it a lot easier to harmonize. Final guidance on the collection and preparation of the medical data was prepared by WP3, task 3.1 and is available in the Data Collection Guidelines which were shared with Data Providers. Furthermore, templates for prospective and retrospective data were developed and distributed to Data Providers to harmonize the data in the INCISIVE Repository. For the pilot study data collection, all guidelines from the retrospective stage, refined and improved were used as well. Then, all the improvements made in pilot study related to Excel templates were corrected and applied for retrospective data to have a homogenized repository.

The challenge which INCISIVE tried 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 of the Project, the responsible Partners undertook annotation of the raw data, as well as produced results, in order to enhance standardization, explainability and findability of this data. The main aim of this processes was to be performed in an automated and standardized way, to the extent possible, minimizing the human intervention. The annotation and labelling of images consisted in designing methods of automatic addition of information to the provided data (ex. tumour location), made available in the INCISIVE Repository, to allow this data to be used as training data for machine learning research. The outputs of these tasks framed 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 consisted 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

Details of the collected metadata for other datasets are described in Annex 1. For example:

- for system telemetry data from INCISIVE applications, the collected metadata consists of application/execution ID, system configuration and additional execution environment data;
- for algorithms performance data, metadata consists 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 were instructed to 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 needed to conform to specific templates.

- Recommended document naming convention was 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 was 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 was to include a version number and date.
- When multiple contributors need to work on a document, it was recommended to use online documents that allow synchronous co-editing.

The research data which had been 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

One of the main outcomes of the project is pan-European repository of medical images (INCISIVE Repository), which enables reuse of data beyond the lifetime of the INCISIVE project, in line with the legal restrictions and ethical guidelines. The Repository is a hybrid 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. Repository is fundamentally a data sharing platform with public access as one of its foundations.

The INCISIVE Repository consists of 9 nodes storing Data from 9 INCISIVE Data Providers. During the final months of the project, INCISIVE designed an action plan to ensure that the Data will be shared post project in line with the legal and ethical requirements. The details are described in Section 7 below. Moreover, the consortium produced a detailed deployment strategy and operational roadmap for the INCISIVE Repository (T8.4).

In parallel, the consortium designed data donorship schema that enables future providers to contribute with their data to the INCISIVE Repository. The mechanism also allows the providers to easily connect their data sources to the INCISIVE Repository. The design of the data sharing schema has been finalised and reported in D5.2 while the implementation has been finalised and delivered as part of the final INCISIVE prototype (D5.3). To support potential INCISIVE data sharing users a dedicated portal was developed to share important information to interested parties and enable them to request access to the INCISIVE platform (available on https://share.incisive-project.eu/).

2) INCISIVE needs assessment study data (WP2)

The results of the WP2 surveys were a part of a report made available to the INCISIVE research team at KU and the Consortium Partners. KU disseminated the findings of the study via journal articles and at relevant research conferences. A summary of the results is available to any participants who request it. It is not possible to identify participants from any such publications, as results are anonymous (unidentifiable) and aggregated for the whole participants' group. For further details please see Annex 1.

3) Other

INCISIVE was committed to make all public academic papers open and free to download. Updated list of publications is available on INCISIVE webpage and will be provided in D9.6.

In public deliverables all personal data was anonymized. Within the limits of privacy laws and intellectual property protection, the digital research data generated in the action will be deposited in public repositories or made publicly available in accordance with the Horizon 2020 Open Access policy.

In particular:

- All public deliverables and open access articles are available on INCISIVE webpage.
- The outcomes produced during UX design workshops with participation of INCISIVE stakeholders (T2.2) included conducting two Delphi studies: one with the primary users of the proposed INCISIVE AI toolbox (healthcare professionals involved in cancer care) and the other with the primary users of the INCISIVE repository. Related deliverable (D2.2) was submitted on 30 September 2021. KU disseminated the findings of the study via journal articles and at relevant research conferences if applicable. See Annex 1 for details.
- AI algorithms and analysis of their performance for semantic segmentation of medical images, disease detection and classification - Models developed within tasks: Prognostics models via spatio-temporal simulation of disease progression, Transforming medical images to medical reports and recommendation, AR enhanced visualization and re-training framework and MaaS and Hybrid learning are integrated in the final INCISIVE platform.

4.3 Making data interoperable

1) Medical images and health data

In early stages of the project, for retrospective data, the Partners focused 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.

In the later stage of the project, the consortium focused on interoperability of the INCISIVE Repository and its compliance with well-established standards, identified in Section 3 above, enabling its communication with the existing systems. Under specific task within the Project, the responsible Partners transformed the data into an interoperable format Common Data Model (CDM). Specifically, the standard on the risk management in medical devices (ISO 14971) as well as IEC 62304 (Medical device software — Software life cycle processes) was considered, while HL7 FHIR and DICOM were followed for interoperability and data exchange between INCISIVE and external systems. Finally, SNOMED and LOINC were used for semantic codification. This further supports interoperability of the data in the INCISIVE Repository with other efforts in this area.

The entire Common Data Model, including the INCISIVE HL7 FHIR implementation, the clinical data semantic codification with SNOMED and LOINC, and the DICOM and NIFTI standardization process are explained in the INCISIVE Interoperability Framework.

Also, the D3.4 INCISIVE Standardization Suggestions provided a summary focused on other projects and external entities on how the CDM made at INCISIVE can be reused and what is the interoperability methodology that INCISIVE recommends after its experience during the project.

1) INCISIVE needs assessment study data (WP2)

Not applicable.

2) Other

As for other data, the (meta)data that was made open and re-usable in line with most widely used terminologies, standards, and methodologies to facilitate interoperability is listed in Annex 1.

From a practical perspective, standard file formats 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).

Details per dataset are provided in Annex 1.

4.4 Increase data re-use (through clarifying licences)

1) Medical images and health data

The conditions of re-use of the INCISIVE Data, including the involvement of Data Access Committee, defining policies on the access to INCISIVE data and Data Use Terms were agreed by the Data Providers. The details are reported under D7.4 IPR Management report.

Furthermore, also from the IP ownership perspective, details on licensing of the data by external data providers in the INCISIVE Repository are defined in Data Donorship Legal Framework (WP7, T7.4) which form legal guidelines and terms, enabling external parties to safely donate their data in the INCISIVE.

2) INCISIVE needs assessment study data (WP2)

N/A

3) Other

Models developed within tasks: Models developed in T4.1, T4.2 and prognostics models via spatio-temporal simulation of disease progression (T4.3), Transforming medical images to medical reports and recommendation (T4.4), AR enhanced visualization and re-training framework (T4.7) and MaaS and Hybrid learning (T4.8) have been integrated in the final platform as part of AI toolbox.

All public deliverables and other public material (publications, communication material, etc.) have been made accessible through the Project web site.

Further details on the licenses are reported in Annex 1 and in D7.4 (in relation to IPR).

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): was 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): was 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): was 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 provided 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 INCISIVE 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.

As described in the Description of Action, the Project results have been mostly published 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. They have been made available by each respective partner's budget and used, where relevant. The details of those publications are provided in appropriate deliverable (D9.6).

5.3 Cooperation with EUCAIM Project

During the project, INCISIVE investigated costs related to the sustainability of the INCISIVE Repository in the post-Project period and possible ways to cover for these costs (business models) was part of the WP8 activities and is discussed in WP8 deliverables.

In particular, INCISIVE acknowledged the launch of EUCAIM funded as a flagship project by the Digital Europe Programme, aiming at the deployment of a ground-breaking federated infrastructure that will power up imaging and AI towards precision medicine for Europe’s cancer patients and citizens. EUCAIM is the cornerstone of the European Commission initiated European Cancer Imaging Initiative, a flagship of the Europe's Beating Cancer Plan (EBCP), which aims to foster innovation and deployment of digital technologies in cancer treatment and care, to achieve more precise and faster clinical decision-making, diagnostics, treatments and predictive medicine for cancer patients. The EUCAIM builds upon the results of the work of the ‘AI for Health Imaging’ (AI4HI) Network which consists of 5 large EU-funded projects on big data and Artificial Intelligence in cancer imaging: Chaimeleon, EuCanImage, ProCancer-I, INCISIVE and Primage. Among other objectives, EUCAIM aims at putting in place and fully deploying all technical and operational measures required for enabling reuse of cancer image data and accompanying clinical data from millions of patients. EUCAIM also aims at enabling the interoperability at technical and semantic level between several existing data infrastructures, including the AI4HI repositories, with INCISIVE being one of the AI4HI repositories.

Given the above scope and the fact that EUCAIM can significantly help in the sustainability and enhancement of the INCISIVE Repository, as well as with the obligation of the INCISIVE Project to make the Data in the INCISIVE Repository available even after the project’s completion, INCISIVE started discussions on facilitating the sharing of its Data through the EUCAIM pan-European digital federated infrastructure of FAIR pan-cancer images. Actions related to the facilitation of the sharing of the INCISIVE Data through the EUCAIM infrastructure may include integration of EUCAIM secure technologies and tools into the INCISIVE infrastructure to ensure technical and semantic interoperability of the INCISIVE Repository with other existing data infrastructures. They may also include adoption of EUCAIM processes and collaboration with EUCAIM bodies that will facilitate the secure sharing of the INCISIVE data through the EUCAIM infrastructure (e.g. integration of EUCAIM’s authentication tools, alignment with EUCAIM’s data request approval processes, integration of improved data security or data interoperability tools that offer equivalent or improved functionality compared to INCISIVE, etc.),

The details on the agreement with EUCAIM will be reported in D7.4.

6 Data security

6.1 Medical images and health data

Initially, the retrospective data required for the development and training of the AI Tool was stored in an infrastructure provided by FTSS IT facilities in Barcelona, Catalonia, Spain, EU (Temporary Infrastructure). The Temporary Infrastructure solution was 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 was in full compliance with various industry and regulatory mandates and ready for audit when necessary. The secure enterprise data hub (EDH) was one in which all data, both data-at-rest and data-in-transit, was encrypted and the key management system was 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 were well trained, an expert certified security procedures and the cluster can pass technical review.

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

The technical Project coordinator with the assistance of Partners with security expertise 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 was already providing data storage and data management services to the Catalonian Health System.

FTSS securely stored the data and made it available to Consortium Partners for their research work during the first months of the Project. This solution was deployed until the final INCISIVE solution for the storage of medical images and data was developed.

After the deployment of the final INCISIVE Repository developed within WP5, the images and accompanying data were stored there. The security features of the INCISIVE Repository are explained in D7.2 Appendix 16: Description of the INCISIVE Repository and Technical and Organisational Measures.

In the following tables, you will find a brief list of security, organizational, and infrastructure measures as they relate to the secure management of the data. The measures are split into general security objectives, to illustrate how these measures come together to ensure the compliant storage, confidentiality, integrity, and availability of the data managed by INCISIVE.

Security Measures

Regulatory Compliance	Preserve Availability
Authentication & Authorization, using Trusted Execution Environment	Data Redundancy configuration applied to Central Node. Guidelines disseminated to Data Providers
Data de-identification, quality checking & curation tools	Automated snapshotting tools deployed Central Node. Guidelines disseminated to Data Providers
Established data provenance procedures and protocols	Distributed storage and service-level infrastructure using NFS and virtualization methods.
Established right-of-removal mechanisms and procedures available.	Disaster Recovery plan in place.
Data minimization & Anonymization protocols defined and applied.	Service-level data-related components are continuously monitored and integrated with automated recovery measures.
	Infrastructure-level components are continuously monitored and integrated with automated recovery measures.
Maintain data Integrity	**Enforce Confidentiality**
Infrastructure auditing & logging tools deployed. Guidelines disseminated to Data Providers	Robust role-based access control measures applied for infrastructure and service resources.
Blockchain transaction tracking & auditing tool deployed	Enforced IAM authorization for infrastructure and service access
Automated backup deployed Central Node; Guidelines disseminated to Data Providers	Restrictive policy for file transfers, enforcing data minimization
Restricted access control measures applied to Central Node. Guidelines disseminated to Data Providers	Data encryption during rare transfer cases (Data provider joining Central Node)

Table 4: Brief list of Security measures

Infrastructure Tools & Practices
Automated service component vulnerability scanning
Service component isolation
Established procedures and schedules for software updating
Established procedures for regular security assessments
Secure Firewall & DNS configurations applied to Central Node. Guidelines disseminated to Data Providers
Procedures & tools deployed for continuous monitoring of infrastructure
ISO-27001 (GDPR) certified infrastructure

Table 5: Additional Infrastructure Tools & Practices

6.2 INCISIVE needs assessment study data (WP2)

All data was entered, stored and backed-up in a secure manner by the research associate for the INCISIVE Project at Kingston University. Once the study was finalised, all personal/identifiable information were 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.

6.3 Other

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

Other research datasets were stored locally, by each relevant Consortium Partner according to their internal rules. Indicative examples follow:

WP2

- UX design workshop data (Task 2.2) 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;

WP3

- Algorithm and methodology for data harmonization (T3.1). The result is the development of a methodology and algorithms that receives the heterogenous data from different sources and integrates them in a specific homogeneous structure and produces error reports. The methodology followed for data integration is already available to the public through a publication. The algorithm was shared with all Data Providers to be used during the project's lifetime. The algorithm will be usable by other parties through a web app that will be integrated in the INCISIVE infrastructure, before uploading the data. The description was available within the related publications (Annex 1)
- Software and data related to the technical requirements of the technology providers, container configuration files to set up the software integration environment (T3.2). This data was stored within a token-secured artifact repository (Azure CR). The development plans were shared among consortium partners, container configuration files however could not be shared with partners or users due to security concerns and because these are project-specific configurations that cannot be generalised and are thus not useful to the wider community.
- System telemetry and performance data from INCISIVE applications (Task 3.4/3.5) was 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;

WP4

- AI algorithms and analysis of their performance data (Task 3.5/4.8) were 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;

WP5

- Metadata about the records of actions of the users on the data in the repository (Task 5.5.) Only authorised members in the platform could check the logs (administrators, medical personnel, organization administrators). Collected personal details include the user's email (the one used to register the INCISIVE platform), the organization they belong to, and their role in the platform. Before joining the platform, users are informed about the use of blockchain and what type of data are collected and used (Privacy Policy).

WP8

- Additional open-source and proprietary market intelligence related data for performing a thorough AI for Health Imaging Market Analysis (T8.1). The market intelligence reports, and other proprietary databases were only be shared within the consortium due to MDT's restraints to not make this information public as per agreed when contracting the services of the business and market intelligence providers. Moreover, this data provides little to no value to any stakeholder group (HCPs, researchers, etc) if used with no specific purpose (i.e. the development of the sustainability strategy and business models) outside the industry of medical devices, in which case it is a strong source of competitive advantage.

- Interviews and/or surveys aimed to validate and improve the INCISIVE service value propositions and business models as well as to establish direct contacts and relationships that may result in a more pro-active adoption of the Project's main exploitable assets (i.e., the AI Toolbox and the Repository) by target groups. All subjects were compiled in a respective contact database (T8.2). The data can be shared solely among project partners, through the project's SharePoint folder and the use of emails. The data shall not be shared outside the consortium because they contain business related information, as well as sensitive personal data (e.g. personal beliefs). All the WP8 deliverables are confidential and as such it will not be possible to share the data contained in sections of the deliverables.

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

- Follow the agreed de-identification 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.

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 had been assessed and required actions were implemented under WP7, which sets out the ethics requirements the Project must comply with. More specifically under T7.2 the required ethical approval actions have been prepared and performed by clinical Partners with the help of the appropriate authorities and other Partners when necessary. The composition of the formal ethics letters, as well as the process for obtaining the formal approvals took place within this task.

Additionally, the Project partners confirmed 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 was not collected beyond the scope of the processing objectives and will not be stored for longer than necessary.

1) Medical images and health data

a) Legal requirements during the Project lifetime

In the context of the Project, the ethical aspects related primarily to the collection and use of medical images and health data and interviews with patients and HCPs. These aspects were addressed on several levels.

Ethics approvals obtained by Data Providers (AUTH, DISBA, UNITOV, UNS, HCS, UoA, VIS, IDIBAPS and GOC) for the collection of retrospective and prospective data were reported under D7.2

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 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 and prospective data were bound by data sharing agreement which sets out their respective tasks and obligations as considered joint controllers of this data. During the terms of the project all the Consortium Partners have signed the data sharing agreements (as enclosed in Annex 2 of the initial DMP). Due to evolution of the project and moving to a different storage infrastructure, the legal framework required amendment. First amendment of agreements was needed due to a new partner, ADAPTIT, joining the consortium and further to BSC's request to copy certain anonymized retrospective data to their supercomputing infrastructure. Second amendment to the data sharing agreement was prepared, consulted by the partners and signed in January 2023. The Amendment 2 addresses: (i) Sharing of data in the INCISIVE Repository, (ii) Use of central node, (iii) Use of data for Inference services, (iv) Legal changes required due to publication of new standard contractual clauses for transferring of data outside of EEA.

The consortium also performed several iterations of Data Protection Impact Assessment (DPIA), which led to identification of additional actions to support data protection compliance and improve data security.

Detailed explanations about the use of retrospective health data in INCISIVE which were resolved in the first year of the project, were explained in D7.1.

b) Legal considerations for INCISIVE to re-share data with third parties post Project

In the context of post-project Data re-use, INCISIVE prepared a detailed plan for the sharing of the Data with external researchers in line with the legal and ethical requirements. The plan included the following requirements:

1. Anonymization of the Data to be re-shared. If the data would remain pseudonymized, this would be contrary to the consents obtained from the patients and the ethics approvals (for prospective data). Also, the data would be considered personal data and GDPR would still apply, making it difficult to provide the data to other researchers (i.e. there would need to be a legal basis for making this data available, for example specific consent). Some data protection authorities do not accept general consent given upfront to broader research use. Also, the dataset would need to be curated, allowing the patients to exercise their rights.
2. The Data would be made available through the INCISIVE Repository only for use and purposes promised to the patients in the consent wording.
3. The project would define the conditions of re-use of the data by the external users, including INCISIVE partners for post project work. After extensive discussions and consultations, these conditions were defined in INCISIVE Repository Data Access Committee, Data Access Policy and Data Use Terms and include in particular that:
    - to access the Data the potential user needs to not only register to the Platform, but apply for obtaining access, in accordance with a defined access procedure;
    - by design, the Data can only be used in the secure environment of the Platform, without downloading or copying the data to external locations;
    - the primary objective of the Platform is to make the Data available for AI training, including Federated Learning (FL);
    - INCISIVE Platform registers Data use transactions, providing an extra layer of accountability for the Data use.

4. Where needed, the Data Providers obtained extension or amendment to existing ethics approvals for retrospective data anonymization.
5. Additionally, the Data Providers were approached to confirm any further (local) requirements which are applicable to them.
6. The consortium agreed on terms of data sharing agreement to be concluded between INCISIVE Data Providers and other consortium partners and the EUCAIM Project. For Further details see point 5.3 above.

   2) Other

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. The relevant Data Providers obtained ethical approvals for pilot studies in each country within T6.4.

7.2 De-identification of medical data in INCISIVE

In the context of medical data the differences between types of de-identification of personal data foreseen by GDPR – pseudonymization and anonymization – become important to understand.

a) Pseudonymization, in GDPR terms, is the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or identifiable natural person (Article 4(5) GDPR). More simply, pseudonymization means that an individual (person) cannot be re-identified based on the pseudonymised data without additional, separate information (key).

b) Anonymization means that an individual (person) cannot be re-identified with reasonable effort based on the data provided or by combining the data with additional data points. More specifically, as GDPR explains, anonymous information is one which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly (recital 26 GDPR).

Anonymization has also other advantages, such as supporting the data minimization principle (Article 5.1.c) GDPR). Still, anonymization of data does not release the controller from all legal obligations. For instance, the controller should ensure that the data remains anonymous, by implementation of ‘appropriate safeguards’.

7.3 Pseudonymization of medical data during INCISIVE project

Given the above, a crucial element of the INCISIVE project was adopting a robust approach to data de-identification, taking into account legal and technical standards and requirements. The concepts of anonymization and pseudonymization were explained by the INCISIVE Ethics and legal manager (TLX) and discussed with other Partners from the very beginning of the project and on several occasions, following the privacy-by-design principle.

For the use of data during the Project, a detailed procedure for the pseudonymization process and guidance for the Data Providers was prepared by the responsible Partners, with support of Legal and Ethics manager. This procedure was applied to the Data which could be submitted in pseudonymized way. Protocol for the pseudonymization of DICOM data and guidelines were documented, providing information for different use cases. The protocol addressed the pseudonymization of data in the DICOM format. Any other scans had to be first converted into the DICOM format and then de-identified. Other imaging data that could not be converted in the DICOM format, contained no personal data and direct or indirect identifiers, so no de-identification process was needed. Clinical information from any other data source were extracted after DICOM patient de-identification and were inserted by clinicians into designated Excel templates to ensure that no direct or indirect identifiers are included in data exchange processes. The pseudonymization of the INCISIVE DICOM images was based on the standard for de-identification of DICOM images which was defined in the DICOM Standard PS 3.15 Digital Imaging and Communications in Medicine (DICOM), Part 15: Security and System Management Profiles. The rules described in the afore mentioned document were grouped in a Basic Confidentiality Profile that removed personal data, while providing options for the removal or retention of several information. Some of these options were either fully or partially implemented, while some of them were not implemented at all.

The information mentioned in the Basic Confidentiality Profile were studied one by one and the decision for the appropriate action to be performed was taken considering INCISIVE’s needs, Basic Confidentiality Profile’s recommended actions and by comparing with the actions taken of other well-known databases (e.g. TCIA).

In summary, the pseudonymization, was conducted in the following manner:

- A selected IT tool was used to remove (or replace with fictional value) certain directly and indirectly identifying data and metadata from the images, such as name of the patient, birthday, date of examination, country of residence. Full list of the data which was removed was carefully considered and listed in a protocol agreed with all Partners.

- Patients were given an ID (code) by the relevant Data Provider. The code was a numerical value. All individual data sets relating to a patient/case (imaging, clinical, histopathological) were linked with by this common code, which referred exclusively to the project.

- Each Data Provider kept a log file with the mapping of the patients with the codes, kept locally and secured. This was never asked by any other Partner of the project.

Considering all of the above, a free-to-use tool was selected, and a certain script provided by the tool was modified so that it met INCISIVE's requirements. Also, sensitive data that were not mentioned in the DICOM standard, but used by the tool were also modified to prevent identification of the patient. A workshop, where all Data Providers participated, was held in order to demonstrate the selected tool and its functionalities, so that it would be easier for Data Providers to use the tool. In addition, detailed guidelines and instructive videos were also made for that purpose. The above procedure was documented in a de-identification protocol.

7.4 Anonymization of medical data for post-project re-use

There continue to be pending discussions and jurisprudence on the meaning (and 'threshold') of personal data and hence its anonymization (see: Breyer decision from the European Court of Justice (ECJ), and other recent cases (T-557/20, C-604/22). There is also no defined process that the Data Protection Authorities or courts recommend for the data anonymization. In view of this limitation, to develop a structure for the process which could be used in INCISIVE, we relied on the Guide to basic anonymization published by National Data Protection Authority of Singapore. This Guide was referred to by the EU Spanish Data Protection Authority on its website. The main steps of the process and their application to INCISIVE are explained in more detail below. There are also other potential solutions and processes possible.

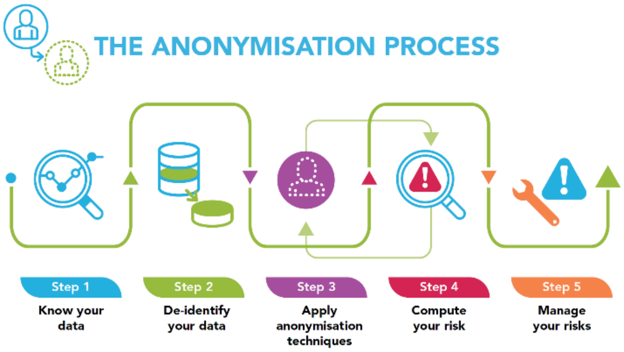


Figure 1: Steps for anonymisation. Source: AEPD. Guide to basic anonymisation. Prepared by the National Data Protection Authority of Singapore (PDPC - Personal Data Protection Commission Singapore).

Step 1: Know your data

In the first step, the Guide asks the controller to evaluate the dataset variables (data fields) by taking into consideration the following criteria: direct identifiers, indirect attributes, and target attributes. This is needed for determination of further steps required to anonymize the data. The Guide states that anonymisation typically involves removal of direct identifiers and modification of indirect identifiers. Target attributes are usually left unchanged, except where the purpose is to create synthetic data.

Step 2: Deidentify your data

The Guide indicates that this step is always performed as part of the anonymisation process. The Guide advises to remove all direct identifiers. Optionally, in this step, it permits to assign a pseudonym to each record if there is a need to link the record back to a unique individual or to the original record. The pseudonyms should be unique for each unique direct identifier. Assignment of pseudonyms should also be robust (i.e. not be reversible by unauthorised parties through guessing or computing the original direct identifier values from the pseudonyms). The pseudonym keys will be deleted in Step 3.

Step 3: Apply anonymization techniques

The Guide requires that in this step the controller applies anonymisation techniques to the indirect identifiers so that they cannot be easily combined with other datasets that may contain additional information to re-identify individuals. Suggested techniques include: • Record or attribute suppression • Character masking • Generalisation • Data perturbation.

The Guide acknowledges that the application of these techniques will modify the data values and may affect utility of the anonymised data for some use cases (e.g. data analytics). After applying the appropriate anonymisation techniques, proceed the controller should proceed to Step 4 to assess the risk level. Then, repeat steps 3 and 4 until they achieve a k-anonymity value indicated below (5 or more).

Step 4: Compute your risk

The Guide refers to k-anonymity method to compute re-identification risk level of a dataset. K-anonymity refers to the smallest number of identical records that can be grouped together in a dataset. The smallest group is usually taken to represent the worst-case scenario in assessing the overall re-identification risk of the dataset. A k-anonymity value of 1 means that the record is unique. The Guidance indicates that the industry threshold for k-anonymity value is at 3 (for internal data sharing) or 5 (for external data sharing). Where possible, a higher k-anonymity threshold value should be set to minimise any re-identification risks. Generally, only indirect identifiers are considered for k-anonymity computation.

Further, the Guide notes that 'if you are not able to anonymise your data further to achieve that, you should put in place more stringent safeguards to ensure that the anonymised data will not be disclosed to unauthorised parties and re-identification risks are mitigated. Alternatively, you may engage experts to provide alternative assessment methods to achieve equivalent re-identification risks'.

All the steps in the anonymization process should be also described and documented.

Step 5: Manage your risk

The Guide states that it is prudent to put in appropriate measures to safeguard data against the risks of re-identification and disclosure. This is in view of future technological advances, as well as unknown datasets that could be used to match against anonymised dataset and allow re-identification to be performed more easily than expected at the time of anonymisation.

As good practice, the details of the anonymisation process, parameters used, and controls should also be clearly recorded for future reference.

It is encouraged to test the anonymization process to detect potential vulnerabilities. Furthermore, periodic security and re-identification reviews during the storage period should be conducted. When integrating data with other datasets, re-evaluation is encouraged, in particular additional layers of anonymization may be applied (e.g. differential privacy or use of synthetic data).

In summary, key points to remember:

a) Anonymization aims to reduce the risk of anonymization below a certain threshold, this is not a binary concept.[6]

b) Throughout the process, data minimisation principle should be followed i.e. only necessary attributes of personal data should be used and shared.

c) It is not enough to remove pseudonyms to render data anonymous.

d) Anonymization cannot be fully automated – tools can help (e.g. to remove direct or indirect identifiers), but the overall context is important as well. Account should be taken of the circumstances under which the data may become identifiable, in particular of the additional information which may be or may become available to the recipient.

e) The less the dataset is modified, the higher the safeguards are needed, and the risk of re-identification is higher.

f) Process of reaching the decision and the assessment of re-identification must be documented.

7.4.1 Anonymization protocol

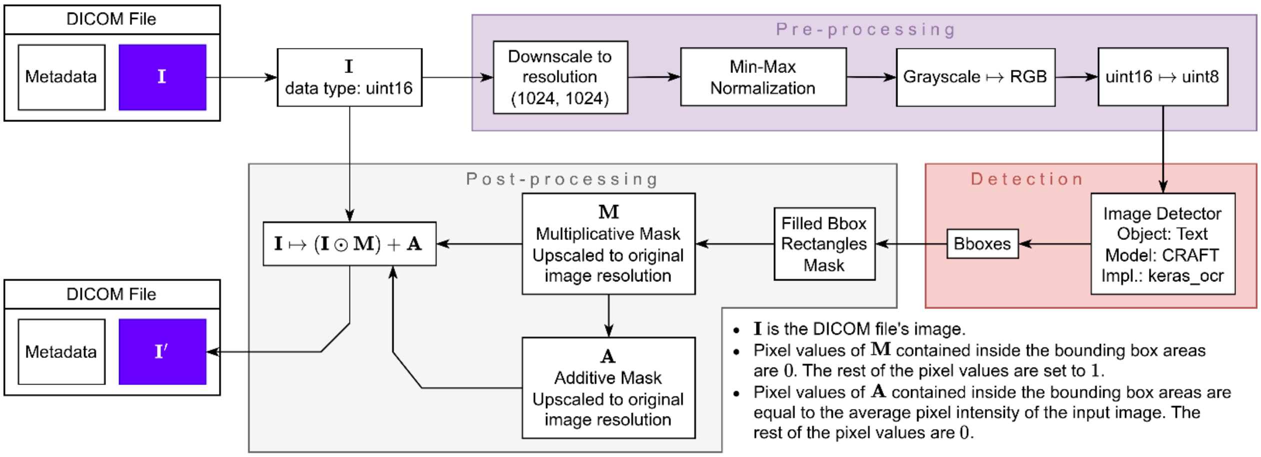
Taking into consideration the above-described process for data anonymization in INCISIVE, the respetive partners structured an anonymization protocol (Step 3: Apply anonymization techniques). It is important to note that Step 1: Know your data and Step 2: De-identify your data, had been achieved previously in the project. The anonymization protocol can be distinguished into three different processes (steps) towards achieving anonymization. These steps included further de-identifying the pseudo-ids of the patients, anonymizing the Excel sheets that contain pseudonymized information, by applying anonymization techniques, such as k-anonymity, and finally using AI algorithms to remove any burned-in text in the DICOM images. The anonymization techniques and their processes are elaborated below:

[6] EDPS: “A robust anonymisation process aims to reduce the re-identification risk below a certain threshold. Such threshold will depend on several factors such as the existing mitigation controls (none in the context of public disclosure), the impact on individuals’ privacy in the event of reidentification, the motives and the capacity of an attacker to re-identify the data” (https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf)

a) **Anonymizing patient pseudo-IDs:** In our effort to ensure the privacy and confidentiality of the medical data within INCISIVE's database, we implemented a robust anonymization protocol. Initially, each data provider was assigned a unique triplet identifier and each patient within a data provider's dataset was allocated a random 6-digit number. These identifiers formed pseudo-IDs, wherein a patient's pseudo-ID comprised the data provider's triplet followed by the patient's 6-digit identifier (e.g., 001-000001). To enhance anonymity, we undertook a comprehensive anonymization process. This involved generating new random prefixes and suffixes for all patients, thereby dissociating them from their original data providers. For instance, previously, a data provider may have been associated with the prefix 001, but post-anonymization, it could be assigned any prefix between 001 and 999. Concurrently, the 6-digit patient identifiers were shuffled to create new suffixes. Following the generation of new IDs, we established mappings between the old pseudo-IDs and the newly anonymized IDs for both data providers and patients. These mappings facilitated the seamless transition from the original pseudo-IDs to the anonymized counterparts, ensuring continuity and accuracy within the database. Subsequently, adjustments were made to the database structure. All patient folders and study folders were renamed based on the newly generated IDs. For example, a folder previously named '001-000001' was renamed to '286-574834', where '286' represents the new prefix and '574834' signifies the new suffix. Similarly, study names within these folders were updated accordingly to maintain consistency. Furthermore, as the pseudo-IDs were embedded in the metadata of DICOM files within each patient's study, modifications were applied to these files. The specific DICOM tags containing the pseudo-ID information were identified and updated according to the generated mappings. This meticulous process ensured that the DICOM files were synchronized with the folder and study name changes, preserving data integrity. Finally, it's imperative to note that the files produced for the needs of anonymization, including the mappings and related information, are temporary and slated for deletion upon completion of the anonymization process. This measure underscores our commitment to data security and privacy, safeguarding sensitive medical information effectively.

b) **Anonymizing the Excel files:** In our commitment to the privacy and confidentiality of medical data for the INCISIVE project, a specialized anonymization protocol for Excel data related to breast, lung, colorectal, and prostate cancer was developed. During the anonymization process, differential privacy was applied to numerical data, ensuring that individual patient data remains confidential while preserving the overall utility of the dataset. This anonymization technique adds a specific amount of noise to the data, and then this noise is carefully calibrated to maintain the utility of the data while protecting the privacy of individuals. The process of applying differential privacy involves selecting an appropriate statistical method for adding noise, such as the Laplace mechanism, based on the sensitivity of the query function and the desired level of privacy. In the Excel sheets, there are some data entries that follow well known standards and terminologies, more specifically the 'who.cc' and 'who.int' terminology for pharmaceutical data. The process of anonymizing those fields included their modification by removing specific terms, aiming to create more generalized terms that still convey the essential meaning without disclosing any details that could be used for re-identification of patients. In case the data in these fields do not follow the appropriate terminology and rather include free text, it is being removed to avoid patient identification. Another key component of our approach involves categorizing patient ethnicities into broader groups. This procedure involves categorizing similar data points into the same clusters, effectively masking individual entries by their common characteristics. Also, the anonymization process includes anonymizing identifiers for both patients and data providers, by replacing them with the newly created IDs that resulted from the process of anonymizing the patient pseudo-IDs (as described in the previous technique). This process not only obscures direct identifiers but also minimizes the risk of re-identification through indirect means.

c) **Image anonymization:**Figure 2: A diagrammatic representation of the Image Anonymization pipeline. A notable proportion of DICOM images contain private patient information in the form of burned-in text (e.g. full name of the patient, date of capture). To ensure the pixel-level privacy of patient data within medical images for the INCISIVE project, we devised an image anonymization pipeline, as shown in Figure 2. Our pipeline effectively eliminates personally identifiable information embedded into the pixel data, and the processed DICOM image data remains conducive to further analysis without compromising sensitive information. Our solution is designed in alignment with EU's GDPR guidelines, as well as NEMA's Attribute Confidentiality Profiles as we add the corresponding DICOM attribute. From a high abstraction perspective, our approach is composed of two steps. The first step involves detecting the permanently embedded text in the pixel data of the image, and the second step concerns the concealment of the predicted text regions in the pixel data. In order to achieve the first step, our implementation is based on CRAFT, a pretrained text detector, we utilized this model through the Keras OCR framework. This model's effectiveness lies in its ability to detect rotated and handwritten text. A wide variety of modalities and scenarios is covered (e.g. MRI/CT/PET including rotated text). For the second step we replace all the detected regions with a solid-coloured layer which inherits the average colour intensity of the image. Following the described process, the header attribute Burned-In Annotation with ID (0028,0301) is added to the DICOM file with a value of 'NO', indicating that the embedded text has been cleaned (as per NEMA's E.3.1).


DICOM File
Metadata
$\mathbf{I}$
$\mathbf{I}$
data type: uint16
Pre-processing
Downscale to resolution (1024, 1024)
Min-Max Normalization
Grayscale ↦ RGB
uint16 ↦ uint8
Detection
Image Detector Object: Text Model: CRAFT Impl.: keras_ocr
Bboxes
Post-processing
Filled Bbox Rectangles Mask
$\mathbf{M}$
Multiplicative Mask Upscaled to original image resolution
$\mathbf{I} \mapsto (\mathbf{I} \odot \mathbf{M}) + \mathbf{A}$
$\mathbf{A}$
Additive Mask Upscaled to original image resolution
DICOM File
Metadata
$\mathbf{I}'$

- $\mathbf{I}$ is the DICOM file's image.
- Pixel values of $\mathbf{M}$ contained inside the bounding box areas are 0. The rest of the pixel values are set to 1.
- Pixel values of $\mathbf{A}$ contained inside the bounding box areas are equal to the average pixel intensity of the input image. The rest of the pixel values are 0.

Figure 2: A diagrammatic representation of the Image Anonymization pipeline.

After a thorough investigation, we have concluded that an optimal resolution for CRAFT's input is 1024x1024. This resolution preserves an appropriate balance between performance and inference time. The developed process complements the previously described anonymization procedures, resulting in a more complete anonymization solution.

After applying these techniques, a testing procedure began to ensure that the applied techniques work properly and produce the expected results. Each time the desired outcomes were not achieved, modifications were made to the anonymization techniques, until the result was satisfying (Step 4: Compute your risk).

7.4.2 Appropriate safeguards under GDPR for post project data re-use

In terms of Step 5, the INCISIVE partners were advised that – also for post project data sharing – appropriate safeguards must be implemented. Those safeguards ensure that the data are not disclosed to unauthorized users and that the re-identification risks are mitigated may include:

- Data can only be shared in the INCISIVE Repository, which has the security measures which are applied in the project (or more stringent).
- Secure data processing infrastructure, including secure central infrastructure, hardware security, robust data transfer protocols, agreed security features of federated nodes, with data breach management and policies implemented by the hosting party.
- Data can only be used in the Repository, for federated training; data cannot be downloaded;
- The use of data is to be tracked /logged through blockchain;
- There are security measures against attacks within the operating system / Hardware Security
- Full data record is not visible to the data users (when use is only allowed for AI federated training). This applies both to the image and the medical data. This is to mitigate the risk

of singling out individuals by the users. The exact scope of data shown vs hidden is not stipulated by the law. It should be determined by balancing the needs of the AI developers (how much data they need to see to perform their research?) and privacy of the patient.

- Access control measures (Identity and Access Management (IAM) system secured with a Trusted Execution Environment (TEE) controls): two factor authentication and authorisation measures restricting the access to data to verified users; regular review of user accounts to ensure all the accounts are active and the rights assigned are necessary
- Access to repository is subject to user verification by data access committee (DAC) or individual DP (for federated data sharing), the potential user has present research proposal for acceptance. Only after the verification, the user can access the repository and use the data.
- DAC or DP should also ensure that the user will not be training an AI model which is malicious (model poisoning can be used to make inferences about the dataset e.g., property or membership inference attacks)
- Data sharing mechanism is in place and includes terms of data sharing, accepted by the DP
- The user must accept the user terms, which include – in particular – prohibition of (i) attempting to re-identify the data subjects, (ii) downloading and coping data outside the repository environment, (iii) circumventing security measures, (iv) modifying original data, (v) use of AI models, which may compromise or copy the data. The terms are defined in INCISIVE Repository Data Access Committee, Data Access Policy and Data Use Terms.

7.5 Data Providers ethical approvals

Ethics approvals were submitted as part of D10.1H - Requirement No. 2 at the end of month 25 of the project. Moreover, the Data Providers DPOs or/and privacy advisors, were contacted to confirm whether the INCISIVE approach to data sharing is acceptable from the local perspective.

7.6 Transfers of personal data within and outside EEA

The status and legal compliance regarding data transfers was reported in D7.2.

7.7 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.

7.8 IPR

Issues regarding the protection of intellectual property rights (IPRs) and confidential information were addressed in detail within the Consortium Agreement. Moreover, existing IP (Background), foreground IP and contributed assets were identified by the Partners in D8.1-D8.5. Assignment of the IPR rights, including agreements regarding sustainability of the Data Repository, will be reported in D7.4 IPR Management Report.

8 Conclusions

The document presents the final INCISIVE Data Management Plan based on the datasets identified by all Partners. The initial draft of the DMP was prepared in M6 of the Project, and this draft was updated and revised two more times during the duration of the Project. The current DMP has been fine-tuned to the data generated during the project and its uses identified by the consortium. The current version also puts focus on making the collected data FAIR and available for re-use beyond the Project term.

9 Annexes

1. Responses to DMP questionnaire