IMI2 Project 101005122 - DRAGON

The RapiD and SecuRe AI enhAnced DiaGnosis, Precision Medicine and Patient EmpOwerment Centered Decision Support System for Coronavirus PaNdemics

WP7 – Technical development of federated machine learning system

D7.1 Data management plan

Lead contributor	P7 Oncoradiomics (ONCO)
Other contributors	Reviewed by the DRAGON consortium
Deliverable submission date	07/04/2021
Deliverable type	ODRP: Open research data pilot
Dissemination level	Public

Abstract

A dedicated Data management plan has been developed as part of WP7 to define the ways data will be used, obtained and stored within the DRAGON project, specifically as it pertains to the distributed learning network developed in WP7.

Methods

The data management plan has been developed based on the H2020 data management plan (DMP) template.

Results

Please see the Data management plan attached at the end of this report. The Data Management Plan developed below addresses the purpose and description of data handled within the DRAGON project and the model for data handling during and after the project, including provisions for data collection, secure long-term storage, integration and interoperability, accessibility and exploitation, in compliance with the principles for findable, accessible, interoperable and reusable research data.

Discussion

This first version of the DRAGON project Data management plan focuses heavily on the imaging and non-imaging data that will be included in the Distributed learning network developed as part of WP7. In future versions of this documents due in month 18 and 36, other types of data produced in this project will be addressed in more detail.

Conclusion

A dedicated Data management plan has been developed specifically for the purpose of DRAGON. It will be updated if needed during the implementation of the project.

Project[1] Number: 101005122

Project Acronym: DRAGON

Project title: The RapiD and SecuRe AI enhAnced DiaGnosis, Precision Medicine and Patient EmpOwerment Centered Decision Support System for Coronavirus PaNdemics

DATA MANAGEMENT PLAN

Contents

Acronyms 3
1. Data Summary 4
2. FAIR data 4
2.1 Making data findable, including provisions for metadata 5
2.2 Making data openly accessible 6
2.3 Making data interoperable 7
2.4 Increase data re-use (through clarifying licences) 7
3. Allocation of resources 7
4. Data security 7
5. Ethical aspects 8
6. Other issues 8
7. Annex 1 – Data Harmonization Specifications 8

Acronyms

AI	Artificial Intelligence
CNN	Convolutional Neural Networks
COVID-19	Coronavirus Disease 2019
CSV	Character-Separated Values
DICOM	Digital Imaging and Communication in Medicine
DMP	Data Management Plan
DRAGON	The RapiD and SecuRe AI enhAnced DiaGnosis, Precision Medicine and Patient EmpOwerment Centered Decision Support System for Coronavirus PaNdemics
FAIR	Findable, Accessible, Interoperable and Reusable
HDF5	Hierarchical Data Format version 5
JSON	JavaScript Object Notation
LAN	Local Area Network
MP4	MPEG-4 Part 14
SVM	Support Vector Machines
WP	Work Package

1. Data Summary

Developing scalable diagnostic and prognostic Artificial Intelligence (AI)-based models requires the use of significant amounts of data. To develop generalisable and efficient models, the data provided during model training should be as diverse as possible. For this reason, the DRAGON project aims at including various data formats and emphasises the deployment of a federated learning system to muster data previously inaccessible for patient privacy reasons.

As DRAGON is aimed at fighting the current coronavirus (COVID-19) pandemic, the data collection effort focuses on gathering clinical data that could help develop diagnostic and prognostic AI-based models to help the treatment of COVID-19. This clinical data originates from the different clinical partners participating to the DRAGON project, as described in Figure 1. Such clinical data includes tabulated data as well as imaging data. Tabulated data is composed of character-separated values (CSV) files adhering to the CDISC standards while imaging data consists in files following the Digital Imaging and Communication in Medicine (DICOM) format.

To develop AI-based models, the clinical data collected by DRAGON will be processed and curated by several software tools. Apart from the AI models themselves (see below), such software tools will generate intermediary data. Depending on the tool used, intermediary data will be stored in various preservation-ready and machine-actionable formats, such as CSV, JSON or HDF5.

The project end-goal – make use of AI to better diagnose and prognose COVID-19 – will result in the generation of several AI-based models. Depending on its type – regressions, support vector machines (SVM), random forests, convolutional neural networks (CNN), each model will be stored in a dedicated, state-of-the-art, format such as HDF5.

DRAGON partner ELF is producing video recordings of patients giving their perspective on their experience with COVID-19 pandemic. These recordings will be produced within the DRAGON project. ELF will obtain informed consent from the patients recoded. This consent will specifically transfer full ownership of the videos to ELF and state that these videos can be made publicly available. All informed consent forms will be kept on file by partner ELF.

Table 1 Preliminary list of data to be evaluated in DRAGON

Type of data	Data format	Number of patients	Modality
Imaging	DICOM	>1000	CT
Clinical data	CSV CDISC SDTM	>1000	COVID-19 diagnosis (PCR), survival, ICU admission, co-morbidities, patient demographics
Video	MP4	3	/

2. FAIR data

Findable, Accessible, Interoperable and Reusable (FAIR) data principles are at the core of DRAGON, as the project aims to federate data repositories of different clinical centres and turn heterogeneous clinical datasets into an interoperable distributed database while ensuring that such distributed database acts as a secure and sustainable platform for data sharing and analysis.

As such, DRAGON data management will involve:

- The development of data and metadata submission standards, controlled terminologies and ontologies as a first harmonisation step between the participating clinical partners. Such harmonisation will ensure the maximal re-use of clinical data to answer future research questions;
- The development of a data pre-processing pipeline acting as a second and final harmonisation layer, as well as a quality check step enabling the training of AI-based models in a distributed setup;

- The development of protocols and standards for intermediary data and metadata storage in dedicated secured repositories to ensure the traceability of every trained model;

- The development of blockchain-based federated learning protocols and infrastructure to allow for a secure and privacy-friendly distributed AI-based model training network.

2.1 Making data findable, including provisions for metadata

Data collection within DRAGON will involve the compliance with protocols for clinical data and metadata submission (formats, standards, etc.) defined in Harmonized Data Specifications (see Annex 1). Each clinical partner should respect these specifications when uploading data to their respective CDistriM workstation. The proposed data submission protocols ensure that the data collected is discoverable, identifiable, and locatable thanks to a standard identification mechanism.

The naming conventions regarding the provided DICOM images should follow the "three-tiered" folder structure and naming convention represented in the following diagram.

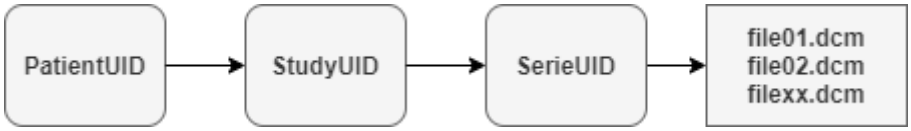

Figure 1: Diagram representing the naming convention for raw data

In this format, the dataset will be stored in one folder that contains a separate folder for each individual patient (that should be named as the PatientUID). In this folder, a separate folder should be created for each different study performed on the patient (that should be named as the StudyInstanceUID). Finally, each study folder should contain a folder for each image acquisition performed on the patient (that should be named as the SeriesInstanceUID) where the DICOM files will be stored with a proper file extension (".dcm").

The naming convention regarding the intermediary files naming convention (how to identify scan, mask and features) will follow this structure:

Figure 2: Diagram representing the naming convention for intermediary data

The scan and mask images will be stored in the Numpy format and their respective metadata will be stored in the Json format. Finally, the features produced will be stored in the Json format. All those files will be named with an identifier generated by the database management system and based on the checksum of each file. Also, versions of those files will be stored in the database directly.

AI model weights will be stored using the HDF5 format using a similar identifier system as the one used by scan and mask images (*i.e.* based on the file checksums). The identifier will be generated by the database management system.

The clinical tabulated data format should use CSV format (or SAS XPORT if an export to CSV is not possible) and follow the CDISC SDTM standard (at least the following domains: Demographics, Procedures, Laboratory tests, Microbiology Specimen, and if possible: Histopathology). Each partner should provide domain files with the following naming conventions:

{partner_name}_{domain_name}.csv

As data will be provided in a cumulative way and model trainings are going to be performed each time that data is provided from a partner, datasets used to produce the models, as well as the models themselves, will be versioned.

The created metadata is composed of different parts: intermediary results, and results with models. Intermediary results will be created to keep track of how the data is used and is specified in the following diagram. The metadata of the final files will be a CSV file containing each radiomic feature computed for model training. Metadate of trained models will be stored in HDF5 files.

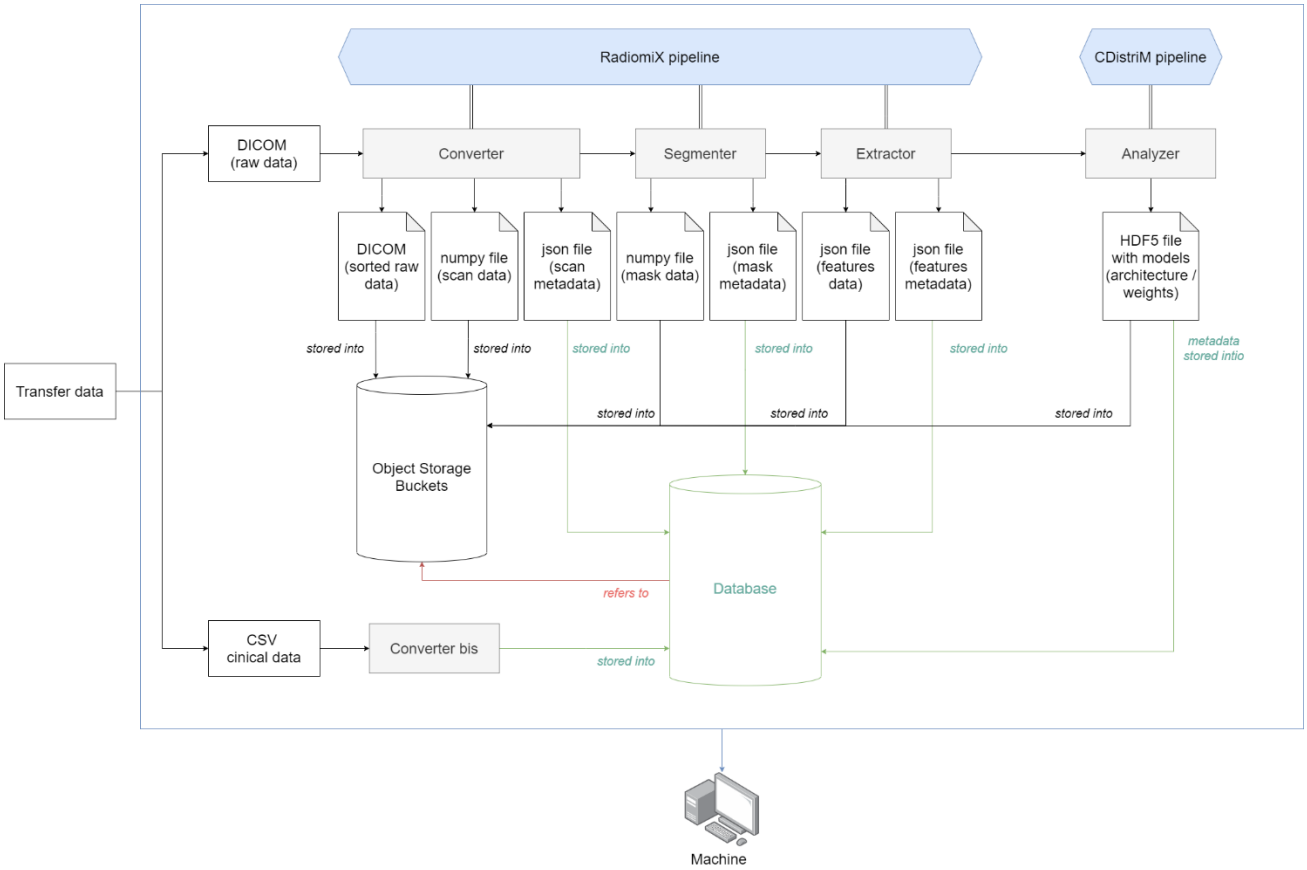

Figure 3: Diagram representing the flow of data during all the process for training and performing models.

As shown in Figure 3, data transferred by partners to their local DRAGON machine will be processed by the RadiomiX pipeline. A first step of quality check will ensure all transferred data is compliant with the Harmonized Data Specifications (see Annex 1). Imaging and other tabulated clinical data will then be processed and stored as described above. Binary data that should not be stored in a database (pixel information, model weights…) will be stored in Object-Oriented Storage buckets while intermediary result metadata will be stored in a local database.

2.2 Making data openly accessible

Open access to research data will be guaranteed as it is defined in the DRAGON grant agreement Article 29.3). Machines supplied by partner Oncoradiomics will be provided to compute radiomics features and train machine learning models by respecting the patient data privacy. In consequence, entry data will not be shared. Each partner can request to integrate models developed for training or validation on the datasets included in the DRAGON distributed learning network, however no one outside of the individual clinical partner will be able to access or view the patient datasets. Final trained model will be transferred to the owner/developer. Moreover, intermediate results may not be used outside of the DRAGON project scope.

All dragon machines will be distributed to the participating centres. Each machine will be accessible by the hosting centre. On top of the hosting centre, only limited set of partners will be granted access to the machines, and only for maintenance purposes (operating system upkeep, software update, new model training code to push to the CDistriM network…).

2.3 Making data interoperable

Data will be interoperable only for the internal use of the different module presented in Figure 3. Data provided by the different partners will not be exchanged between any researchers, institution, organisations, etc. Also, there will be no re-use of the data allowed for any purpose not related to the DRAGON project.

Within each DRAGON machine, all intermediary files represented in Figure 3 will be referenceable by the metadata stored in the local database. This will allow traceability and access management at the database level. Each local database content will be harmonized to CDISC SDTM standards to ease potential data submission to authorities (FDA, CE) in the context of a medical device certification.

The CDISC SDTM standards terminology will be followed for metadata represented in Figure 3 saved in the database. The image data will be harmonized using the Numpy format and the model data (containing the architecture and weights) will follow the HDF5 format.

As the Numpy and HDF5 format are not part of standard libraries, a fixed vocabulary specific to the data type (Numpy or HDF5) that will not follow the CDISC SDTM standard, will be used. Also, the HDF5 model produced will follow a mapping pipeline that will make accessible all the sub-results (model weights, architectures, metrics with all variables, etc) with a common terminology.

2.4 Increase data re-use (through clarifying licences)

DRAGON will provide an infrastructure via a distributed learning network which will enable allow to maximize re-use of COVID-19 patient imaging and clinical data within the ethical and legal framework for responsible data sharing. The data included in the distributed learning network falls under the “secondary use of data” approval by the local ethical committees, the data included here will therefore be re-used and serve an additional purpose than that for which it was originally acquired for.

3. Allocation of resources

DRAGON project aims at including various data formats and emphasises the deployment of a federated learning system to muster data previously inaccessible for patient privacy reasons. As such, making data FAIR is an integral part of the project and is already covered in the various costs of project related to the development and deployment of CDistriM and the design of common data curation processes.

As the bulk of data standardisation effort falls within the scope of work package 7 (Federated Learning), OncoRadiomics is the responsible for data management.

The data included in the distributed learning network will be stored and archived by the clinical partners according to the local data management policies.

Within DRAGON, tasks in work package 11 (WP11) are aimed specifically at ensuring sustainability of the results and infrastructure developed as part of the project. WP11 will work to develop a plan to sustain the collaboration and the resulting assets after the funding period. The longer term impact will be substantially increased if this project leads to a sustained effort that continues to build over time. A comprehensive plan for long term preservation will be developed as part of the project.

4. Data security

Within the federated learning paradigm, patient data never exits the network partner systems. As such, it already provides a safer platform than central approaches by removing the need of various data transfer steps. The DRAGON machines will host the patient data locally and only transfer "abstract representations" of such data in the form of the trained model weights.

Each DRAGON machine will host a copy of the patient data for the duration of the project. This means that the original data will remain safely in the partner database system, and that the copied data will be erased from the DRAGON machine at the end of the project if no secondary use of the data was agreed upon.

The model weights are stored outside of the local machine, at the federation server level, which will be hosted by Google Cloud Services (GCS). GCS provide a state-of-the-art approach to data security, and data privacy, as stated in GCS privacy policy and service level agreement.

To provide an additional layer of security, a "Security by Design" approach was taken when designing the requirements for DRAGON machines. Such approach includes drastically limiting the connectivity of the DRAGON machines to a sub-network of the partner LAN and a single distant IP address (the federation server). Every data communication, and every data storage on the DRAGON machines and on the federation servers will be encrypted using state of the art methods.

5. Ethical aspects

Data manipulated for this project is private and personal data. Data is pseudo-anonymized by the relevant clinical partner by using unique identifiers to ensure data privacy and clinical ethic.

Data provided by partners for this project was acquired as part as previous studies. As such, such data falls in the category of secondary re-use. Partners got the approval of their Steering Committee (or primary investigator or research group) and Medical Ethical Committee to provide data for this project. As such, patient data can only be used for this project and other re-use will be submitted to the same authorisation process.

There is no sharing and no long-term preservation for raw data. In case this policy would change, a written consent from the different partners and patients should be required.

Prior to inclusion of data to the distributed learning network, the partners must check if special derogations pertaining to the rights of data subjects or the processing of genetic, biometric and/or health data have been established under the national legislation of the country where the research takes place and provide a declaration of compliance with respective national legal framework(s).

Detailed information on the specific data protection informed consent procedures in regard to data processing must be provided prior to the start of the research activity. The specific data protection informed consent forms and information sheets (in language and terms intelligible to the participants) are to be kept on file by the clinical partner responsible for data collection and sharing via the distributed learning network.

6. Other issues

We will not make use of other nation, funder, sectorial or departmental procedures for data management.

7. Annex 1 – Data Harmonization Specifications

This document is intended to guide the organization, structure, and format of data to be transferred to a DRAGON machine. Transfer of the data should happen according to the specifications below.

Harmonized Data Specifications

For protocol IMI2 Project 101005122 - DRAGON

Version 1.0

Clos Chanmurly 13 – 4000 Liège – Belgium – +32 4 229 22 80 – www.radiomics.bio

1. General

This document is intended to guide the organization, structure, and format of data to be transferred to a DRAGON machine. Transfer of the data should happen according to the specifications below.

2. Contact Information

Company	Radiomics (Recipient)		(Provider)	
Contact	Benjamin Miraglio			
Role	Head of Technology			
Address	Clos Chanmurly 13			
City, State, Zip	4000 Liège			
Country	Belgium			
E-mail	benjamin.miraglio @radiomics.bio			

3. Transfer Specifications

Schedule	☒ Final ☐ Annual ☐ Quarterly ☐ Per Batch ☐ On request
Anonymised	☐ No ☒ Yes
Transfer Type	☐ Cumulative ☒ Incremental
Data Format	☒ DICOM ☐ Nifti ☐ Numpy ☒ CSV ☐ Other
Medium	☒ SFTP ☐ DICOM ☐ Hard-Drive ☐ Other

4. General Data Requirements

4.1. DICOM files

All data must be curated to contain the DICOM information listed in **Table 1** and **Table 2**.

Table 1 - DICOM File Requirements

DICOM keyword	DICOM Tag	Value type	Requirement
SeriesInstanceUID	(0020,000E)	String	Unique Identifier
StudyInstanceUID	(0020,000D)	String	Unique Identifier
PatientID	(0010,0020)	String	Identifier
Modality	(0008,0060)	Code String	"CT"
SliceThickness	(0018,0050)	Decimal String	Value included in [0.5 - 1.5]

PixelSpacing	(0028,0030)	Decimal String	Value included in [0.6 - 1.5]
ConvolutionKernel	(0018,1210)	Short string	e.g. "SOFT", "STANDARD", "DETAIL", "B", "C", "B30f","B30s","B31f","B31s","B35f", "FC08", "FC24"
Manufacturer	(0008,0070)	Long String	/
ManufacturerModelName	(0008, 1090)	Long String	/
DeviceID	(0018,1003)	Long String	/
SeriesTime	(0008, 0031)	Time	/
SeriesDate	(0008, 0021)	Date	/
SeriesDescription	(0008,103E)	Long String	/
StudyTime	(0008, 0030)	Time	/
StudyDate	(0008, 0020)	Date	/
StudyDescription	(0008,1030)	Long String	/

Table 2 - DICOM tag units

DICOM keyword	DICOM Tag	Value type	Example	Unit
SliceThickness	(0018,0050)	Decimal String	1.5	mm
PixelSpacing	(0028,0030)	Decimal String	[0.644, 0.644]	mm

Also, the organizational structure of provided DICOM files will follow the naming convention as shown in **Error! Reference source not found.**.

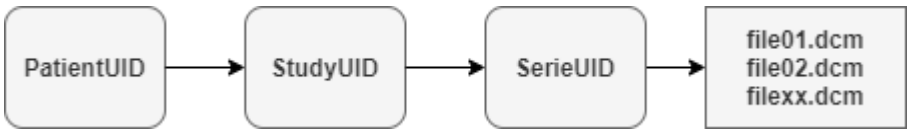

Figure 1 - Diagram representing the naming convention for raw data

A CSV file containing global metadata must be provided using columns defined in **Table 3**. This table gives an overview of Radiomics' request regarding the structure of provided data. All columns are required but the list is not exhaustive and other information can be added such as annotation or else. This document allows the provider and the recipient to make a first quality check on provided data (*e.g.* How many subjects? How many slices per scan? …).

Table 3 - Metadata Structure

Column	Description	Data Type	Length	Examples/Options
project	Project Identifier	Char	20	IMI2 DRAGON
partner	Partner Identifier	Char	10	e.g. UNIFI
patient	Patient Identifier	Char	10	e.g. PATIENT01
modality	Name of the modality	Char	5	CT
date_time	Date and time of the visit	Char	20	e.g. 2020/12/03 13:30:26
study_uid	Study unique identifier	Char	100	
series_uid	Series unique Identifier	Char	100	
series_date_time	Date and time of the series	Char	20	e.g. 2020/12/03 13:30:26
slices_number	Number of slices per series	Num	3	
convolution_kernel	Name of the convolution kernel	Char	10	
slice_thickness	Slice thickness	Num	10	
pixel_spacing	Pixel Spacing	Num	10	

4.2. CSV files

Clinical data will be provided in the CDISC SDTM standard in a CSV format, and it may contain theses different domains as much as possible:

- Demographics (DM), *required*
- Substance Use (SU), *optional*
- Laboratory Test (LB), *required*
- Microbiology Specimen (MB), *required*
- Microscopic Findings (MI), *optional*
- Death Details (DD), *optional*

Each domain file requires at least some variables defined in following tables. The list of variables is not exhaustive and should refer to CDISC SDTM Terminology and SDTM Implementation Guide v3.3 if needed.

Table 4 - Demographics domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
DOMAIN	Domain Abbreviation	Char	DM

USUBJID	Unique subject identifier	Char	
AGE	Age of the subject	Num	
SEX	Sex	Char	F/M
ETHNIC	Ethnicity	Char	HISPANIC/LATINO
RACE	Race	Char	BLACK/NATIVE AMERICA/ALASKA NATIVE
COUNTRY	Country	Char	

Table 5 - Substance Use domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
DOMAIN	Domain Abbreviation	Char	SU
USUBJID	Unique subject identifier	Char	
SUSEQ	Sequence number	Num	
SUTRT	Reported name of substance	Char	Cigarettes/Coffee/ …
SUDOSE	Dose of the substance	Num	
SUDOSU	Dose unit of the substance	Char	CIGARETTE EQUIVALENTS/GRAMS/…
SUDUR	Duration of substance use	Char	(ISO 8601)

Table 6 - Laboratory Test domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
DOMAIN	Domain Abbreviation	Char	LB
USUBJID	Unique subject identifier	Char	
LBSEQ	Sequence number	Num	
LBTESTCD	Test code	Char	
LBTEST	Test name	Char	PLATELETS/ …
LBCAT	Category for lab test	Char	HEMATOLOGY/URINALYSIS/CHEMISTRY/ …
LBORRES	Original result	Char	
LBORRESU	Original result unit	Char	
LBSTRESC	Standard character result	Char	
LBSTRESN	Standard numeric result	Num	

LBSTRESU	Standard result unit	Char	

Table 7 - Microbiology Specimen domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
DOMAIN	Domain Abbreviation	Char	MB
USUBJID	Unique subject identifier	Char	
MBSEQ	Sequence Number	Num	
MBTESTCD	Test code	Char	
MBTEST	Test name	Char	COVID tests (PCR, BAL)
MBTESTDTL	Test or examination detail	Char	PRESENT/COLONY COUNT/…
MBORRES	Original result	Char	
MBORRESU	Original result unit	Char	
MBSTRESC	Standard character result	Char	
MBSTRESN	Standard numeric result	Num	
MBSTRESU	Standard result unit	Char	
MBDTC	Date and time of collection	Char	(ISO 8601)

Table 8 - Microscopic Findings domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
DOMAIN	Domain Abbreviation	Char	MI
USUBJID	Unique subject identifier	Char	
MISEQ	Sequence Number	Num	
MITESTCD	Test code	Char	
MITEST	Test name	Char	Breast Cancer Susceptibility Gene 1/ …
MIORRES	Original result	Char	
MIORRESU	Original result unit	Char	
MISTRESC	Standard character result	Char	
MISTRESN	Standard numeric result	Num	
MISTRESU	Standard result unit	Char	

MISPEC	Specimen material type	Char	

Table 9 - Death Details domain structure

Domain Variable	Description	Type	Example/Option
STUDYID	Study Identifier	Char	
USUBJID	Unique subject identifier	Char	
DDTESTCD	Test code	Char	
DDTEST	Test name	Char	Primary Cause of Death
DDORRES	Original result	Char	
DDSTRESC	Standard character result	Char	
DDDTC	Standard result unit	Char	(ISO 8601)

All domain files need to follow the naming convention bellowing:

{partner_name}_{domain_name}.csv

5. Additional Requirements

As agreed by both parties, and as specified in the Consortium Agreement, the data will be transferred to the Recipient sequentially.

The data shall verify the requirements:

- The data shall not contain patient identifying information;
- Every data belonging to a same patient shall be clearly linked through a common patient identifier or an association table;
- The data shall not contain any missing values;

On top of these requirements, if the data is a DICOM scan, it shall also verify the following additional requirements:

- The scan shall be complete and not contain any missing slices;
- All DICOM files belonging to a same scan shall contain the exact same values for every DICOM tag represented in Table 1;
- The chronological order of two scans belonging to a same patient shall clearly be deductible from the “SeriesDate” DICOM tag or mentioned in an association table;
- The scan shall contain at least 25 slices and no more than 1500 slices;

6. Checks Strategies

In order to provide a clear and unambiguous communication between Radiomics and the different partners in terms of data specification and transmission, Radiomics’ Data Scientist and Manager need to receive data sets in the above specified format. To this end, the following checks will be applied to the data, if applicable:

Checks	**Privacy:** 1. ☒ Check for unexpected identifying personal data. **Uniqueness:** 2. ☒ Check for uniqueness of series instance identifier. **Completeness:** 3. ☒ Check if the correct number of slices are available. 4. ☒ Check if every DICOM tags mentioned in the HDS are available. 5. ☒ Check if for a chosen scan, every DICOM files contain the same value for every DICOM tags. 6. ☒ Check if the chronological of two scans is deductible and coherent. 7. ☒ Check if clinical data contains no missing values. **Accuracy:** 8. ☒ Check if scans are compliant to DICOM standard. 9. ☒ Check if the pixel spacing respects limits defined in the HDS. 10. ☒ Check if the slice thickness respects limits defined in the HDS. 11. ☒ Check if clinical data is respecting CDSIC SDTM standard. 12. ☒ Check if clinical data contains domains mentioned in the HDS.

7. Revision History

Version Number	Date	Reason for changes	Author(s)
1.0	25MAR2021	Initial Document	Adeline Rossetti-Morot Nathan Tsoutzidis

8. Approvals

Benjamin Miraglio, PhD Head of Technology Email: benjamin.miraglio@radiomics.bio	**Signature**
Person, Title Role Email:	**Signature**
Person, Title Role Email:	**Signature**