D3.5

Initial Data Management Plan

Related Work Package	**WP3 – Dissemination, Communication, Open data availability and Data management Plan**
Related Task	Task 3.3 - Data Management Plan and Open Research Data Pilot
Lead Beneficiary	FORTH
Contributing Beneficiaries	All
Document version	vf (final version)
Deliverable Type	Report
Distribution level	Public
Contractual Date of Delivery	31/03/2021
Actual Date of Delivery	31/03/2021

Authors	**Nikolaos Tachos, Eugenia Mylona, Haridimos Kondilakis, Manolis Tsiknakis**
Contributors	Daniele Regge, Nikolaos Papanikolaou, Kostas Marias, Dimitris Fotiadis
Reviewers	Emily Johnson, Max Königseder, Elena Remacha

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement **nº 952159**

Version history

Version	Description	Date completed
0.0	Table of Contents	29.01.2021
0.5	Initial version of the Deliverable	26.02.2021
0.6	Contributions by partner on selected sections	12.03.2021
0.9	Version 1.0 ready for internal review	19.03.2021
1.0 (vf)	Final version incorporating the replies to the review comments	26.03.2021

Statement of Originality

This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Disclaimer

Executive summary

This deliverable is the initial version of the Data Management Plan (DMP) of the Pro-Cancer-I project, in accordance to the regulations of the Pilot action on Open Access to Research Data of the Horizon 2020 program (H2020). It contains provisional information about the data that will be produced and collected within the project, whether and how it will be made accessible for re-use and further exploitation, and how it will be curated and preserved.

Information about the main categories of the ProCancer-I data are provided in the Data introduction section. In Section 2, we provide a detailed description of the data that will be gathered and processed for each work package (WP), as well as of the data collection, documentation, metadata generation and data assessment mechanisms. Furthermore, the tools for data storage and methodologies for ensurying data security are also presented in this section. Next, Section 3 describes how the data management plan is aligned with the FAIR principles. General Data Protection Regulation (GDPR) regulation and its application to the ProCancer-I project is described in section 4, while ProCancer-I Open Data in Section 5 deals with the participation of the ProCancer-I in Open Research Data Pilot (ORDP). Finally, the project risks and proposed mitigation measures and related ethical aspects raised by ProCancer-I research are provided in Section 6 and 7, respectively.

This is an initial version of the Data Management Plan of the ProCancer-I system. The datasets described at this stage, represent an early reflection on the data that we foresee to collect or generate. Given that the DMP is a living document, it is expected to be further modified or detailed during the lifetime of ProCancer-I. The information can be made available on a finer level of granularity through updates as the implementation of the project progresses and when significant changes occur. Those changes might include new data, changes in consortium policies (e.g. new innovation potential, decision to file for a patent) or changes in composition and external factors (e.g. new consortium members joining). Nevertheless, main principles - as described within this deliverable - is expected to remain intact until the end of the project, thus forming the main strategic axes of the overall Data Management Plan.

At minimum, the DMP will be updated in the context of the periodic evaluation/assessment of the program, but it is foreseen that the implementation of the DMP at project level will also be part of the annual reporting. The final version of the DMP will be delivered at M36 within the deliverable “D3.6 Final data management”.

Table of Contents

List of Abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
DICOM	Digital Imaging and Communications in Medicine
DMP	Data Management Plan
DoA	Description of Annex
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable, Re-usable
GDPR	General Data Protection Regulation
IPR	Intellectual Property Rights
MRI	Magnetic Resonance Imaging
mpMRI	multi-parametric Magnetic Resonance Imaging
ORDP	Open Research Data Pilot
PCa	Prostate Cancer
WP	Work Package

List of Tables

List of Figures

1 Introduction

1.1 Purpose of the ProCancer-I Data Management Plan

ProCAncer-I's vision is to become a catalyst in this process by creating the first European, ethical- and GDPR compliant, quality-controlled, prostate related, medical imaging platform, in which both large-scale data and AI algorithms will co-exist. The project will create ProstateNET to be the largest repository worldwide of high-quality mpMRI PCa images. Given that the majority of the ProCancer-I datasets involve data collection from human participants, the respective data produced, raw or processed, will be carefully handled, under thorough consideration of ethical and privacy issues involved in such datasets. For all the identified ProCancer-I datasets, specific parts that can be made publicly available have been identified in the current first version of the project's DMP.

The DMP of ProCancer-I realizes the data management regarding two types of data: on the one hand the utilization of the research data that are generated and collected within the context of the project, and on the other hand the dissemination of the scientific results generated from the project.

The present deliverable is developed based on the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, as well as to the General Data Protection Regulation and is structured taking into account the Horizon 2020 FAIR Data Management Plan[1]. According to the guidelines on Data Management in Horizon 2020, a dataset description template has been drafted to provide the main pillar for the dataset descriptions (Annex I). Relevant datasets have been recognized and a detailed list has been generated, based on the datasets that have been described within the DoA to be produced during the lifetime of the project. The deliverable consolidates all the partners' feedback and provision for the datasets they contribute to.

1.2 Background of the ProCancer-I Data Management Plan

ProCancer-I DMP is in accordance with the following articles of the Grant Agreement (GA):

Article 29.2 Open access to scientific publications

Each beneficiary must ensure open access (free of charge online access for any user) to all peer-reviewed scientific publications relating to its results.

Article 29.3 Open access to research data

Regarding the digital research data generated in the action ('data'), the beneficiaries must:

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

(a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

(i) the data, including associated metadata, needed to validate the results presented in scientific publications, as soon as possible;

(ii) not applicable;

(iii) other data, including associated metadata, as specified and within the deadlines laid down in the 'data management plan';

(b) provide information — via the repository — about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and — where possible — provide the tools and instruments themselves).

Article 36 Confidentiality

During implementation of the action and for four years after the period set out in Article 3, the parties must keep confidential any data, documents or other material (in any form) that is identified as confidential at the time it is disclosed ('confidential information').

Article 39.2 Processing of personal data by the beneficiaries

The beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection (including authorisations or notification requirements).

The beneficiaries may grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement.

The beneficiaries must inform the personnel whose personal data are collected and processed by the Commission. For this purpose, they must provide them with the privacy statement(s) (see above), before transmitting their data to the Commission.

2 Data description

2.1 Data Types

Different data types will be collected during the course of the project, such as publications and research data, managerial and ethical documents as well as anonymized patients' clinical, biological and imaging (MRI) data. A list of the different datasets has been established, and this list will be further detailed to precise the type of data generated. Data collection will rely on both data already available to partners at the beginning of the project and on data that will be generated during the lifetime of the project. Section 2.3.2 provides a summary of the data that will be collected from each WP of the ProCancer-I project.

All the data generated and collected by the project will be stored in a centralized data repository, in order to improve standardization and promote reusability. Data formats will be selected with the view to facilitate data storage and reusability. Therefore, data will be in both human-readable and machine-readable format (e.g. RDF, XLM and JSON). Additionally, when possible, non-proprietary formats will be used. More detailed information regarding the format of the data is provided in section 3 - FAIR data.

2.2 Data users

The data collected and generated through the ProCancer-I project may be exploited by a wide range of data users, including:

- Medical Doctors
- Researchers on PCa
- AI model developers
- Project's partners: Clinical partners, Companies, Research Institutes
- Wider audience
- European Commission

2.3 ProCancer-I Datasets

2.3.1 Datasets naming

The convention followed for naming the project datasets, is the following:

1. A prefix "DS" indicating a dataset.
2. Its unique identification number depending on the WP the dataset comes from, e.g., "DS1" for datasets coming from WP1, "DS2" for datasets coming from WP2 etc.
3. A serial number restarting at 1 for each WP indicating the sub-dataset come from the specific WP: "DS1.1", "DS1.2" etc.
4. A short name indicative of its content and purpose. e.g., "DS1.4_managerial documents".
5. If a versioning of the DS is needed then the latter in placed at the end of the naming. e.g, "DS1.4_managerial documents_v1".

2.3.2 Summary of the datasets

Tables 1-8 present a short description of the content of the ProCancer-I datasets.

***Table 1**: Datasets of WP1.*

Datasets of WP1	The datasets of WP1 contain information related to the project management and coordination.	
DS1.1	Partners contact list	This dataset contains the detailed consortium contact information.
DS1.2	Financial statements	This dataset contains the financial statement log file describing the financial statement reports along with a small description.
DS1.3	Risk log	This dataset contains the identified risks from the beginning of the project accompanied with the mitigation plans.
DS1.4	Managerial documents	This dataset contains a list of the managerial documents that will be prepared within the lifecycle of the project.

***Table 2**: Datasets of WP2*

Datasets of WP2	The datasets of WP2 contain information related to the ethical and legal requirements.	
DS2.1	Ethical Aprovals	This dataset includes the list of the documents related to ethical approvals.
DS2.2	GDPR	This dataset includes the list of the documents related with GDPR issues.

***Table 3**: Dataset of WP3*

Datasets of WP3	The datasets of WP3 contain information related to the project dissemination and exploitation.	
DS3.1	Communication KPIs	This dataset includes the list of the communication Key Performance Indicators (KPIs) (e.g. KPI name, target value, achieved value, plan for the achievement of the target value in case it is needed etc.)
DS3.2	Dissemination materials	This dataset includes the list of the dissemination materials that will be developed within the project lifecycle (e.g. type, content etc.)

DS3.3	Exploitation plan	This dataset will include the list of documents related to the exploitation plan and the exploitation activities performed.
DS3.4	Contact details of linked initiatives	This dataset will include the contact details of linked initiatives that the consortium communicate with during the lifecycle of the project.
DS3.5	IPR	This dataset will include a list with the components of the See Far solution along with the IPRs per partner.

Table 4: Datasets of WP4.

Datasets of WP4		The datasets of WP4 contain information related to the security, privacy, transparency and sharing of the data repositories and the AI models provided by the ProCancer-I platform.
DS4.2	Data representation	This dataset contains a list of ontologies and data models that will be used to ensure machine-actionable data representation.
DS4.3	Standards for safety and privacy	This dataset contain information related to the standards that will be followed to ensure data safety and the requirements for data anonymization.

Table 5: Datasets of WP5.

Datasets of WP5		The datasets of WP5 contain information related to the data that will be collected for the development of the AI models.

DS5.1	Clinical data	A detailed list of patient data with a complete medical history (assessed before the beginning of the program) including:
		• Identification and demographics (name, patient ID, birth date, height, weight etc.)
		• Medical History (including major illnesses, family history)
		• PSA and PSA density (PSAD)
		• Pathological findings:
		  ○ Gleason score
		  ○ Status of resection margins (in case of radical prostatectomy)

		○ Presence of extraprostatic invasion ○ Nodal status • Treatment-related data: ○ Active Surveillance (AS) ○ Treatment type ○ Time to metastasis ○ Time to biochemical recurrence ○ Toxicity data after radiation treatment ○ Quality of Life assessment after treatment ○ Gleason follow-up and/or biopsy for AS
DS5.2	Imaging data	The collected imaging data will consist of prostate mpMRI in DICOM format, including: • T1-weighted sequences • T2-weighted sequences • Diffusion-weighted imaging (DWI) • Dynamic contrast-enhanced (DCE) sequences • Apparent diffusion coefficient maps (ADC) • Annotation marks of the regions of interest. Imaging data coming from Siemens, Philips and GE MRI systems using 1.5T or 3T field strength will be collected from the local PACS system of each clinical parner.

Table 6: Datasets of WP6.

Datasets of WP6	The datasets of WP6 contain information related to ethical, trustworthy and FAIR aspects of AI models.
DS6.1	Performance monitoring	This dataset contains a list of performance reports and measurement benchmarks for the AI models (including old and updated versions).

Table 7: Datasets of WP7.

Datasets of WP7		The datasets of WP7 contain information related to the performance the AI models in the context of 8 clinical scenarios.
DS7.1	Error tracking	This dataset contains the list of errors, failures or innacuaracies occuring during AI-model validation in real-world scenarions.
DS7.2	Safety and effectiveness	This dataset contains a list of performance reports on external data. This dataset can be merged with DS6.1.

Table 8: Datasets of WP8.

Datasets of WP8		The datasets of WP8 contain information related to the sustainability planning and business opportunities.
DS8.1	Sustainability & Business plan	This dataset contains the list of sustainability actions and potential investors. (This dataset can be merged with DS3.3)

2.3.3 Datasets description

In compliance with the template provided in Appendix I – Dataset description template, the Tables 9-16 present in detail the information regarding the ProCancer-I datasets, in terms of: i) generic description, ii) origin of data, iii) nature and scale of data, iv) to whom the dataset could be useful, v) related scientific publications, vi) indicative existing similar data sets, vii) partners activities and responsibilities, viii) standards and metadata, ix) data exploitation and sharing, x) archiving and preservation.

Table 9: DS1.1_Partners Contact List description.

Data identification: *DS1.1_ Partners Contact List*
Generic description:
The contact details of the persons representing each partner organization in the See Far project and participating in each WP and task. Contact details include telephone number, skype name and address.
Origin of data:
The data was collected at the beginning of the project and they will be updated once a change in the personnel of each organization (ProCancer-I partner) takes place.
Nature and scale of data:
Spreadsheet data
To whom the dataset could be useful:
All partners
Related scientific publication(s)
N/A
Indicative existing similar data sets (including possibilities for integration and reuse):

N/A	
Partners activities and responsibilities	
Partner owner of the data	ProCancer-I consortium
Partner in charge of the data analysis	ProCancer-I consortium
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	All
Standards and metadata	
Standards, format, estimated volume of data	Excel format files
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data are involved. All the partners have agreed on this during the kick off meeting of the project.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3 years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 10: *DS1.2_Financial statements description.*

Data identification: *DS1.2_Financial statements description*
Generic description:
The financial information of each partner of the consortium will be included in the DS1.2 dataset. They will include information about the personnel cost, the justification of travel, the justification of Equipment, the justification of other goods and services, the justification of sub-contracting, the justification of linked third parties and the Justification of contributions linked third parties.
Origin of data:
The information will be provided by each partner to the coordinator along with the interim progress report in M6, M12, M18, M24 and M36.

Nature and scale of data:	
Spreadsheet data	
To whom the dataset could be useful:	
All partners, European Commission.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	ProCancer-I consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files.
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3 years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 11: *DS1.3_RiskLog description.*

Data identification: *DS1.3_RiskLog description*
Generic description:

A description of the risk, its causes, the kinds of problems that it could result in (potential effects), and risk dependencies.	
Origin of data:	
Several potential risks have been identified with direct or indirect impact on ProCancer-I solution. Risks are grouped into four categories: a) general and administrative, b) technical and scientific, c) exploitation and dissemination, d) ethical. Each partner will estimate and evaluate the associated risks, the respective controls and will monitor the effectiveness of the controls in collaboration with RM.	
Nature and scale of data:	
Spreadsheet data. The Risk Log for the project is using PM2 Risk Log template and no changes have been done to the structure, fields or values.	
To whom the dataset could be useful:	
Executive Board and Project Core	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	ProCancer-I consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	PM2 Risk Log template
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3 years following the end of the project.

Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 12: *DS1.4_Managerial documents description.*

Data identification: *DS1.4_Managerial documents description*	
Generic description:	
Information related to the project management and coordination.	
Origin of data:	
The information will be collected through the whole lifecycle of the project.	
Nature and scale of data:	
Documents in excel, word and pdf format.	
To whom the dataset could be useful:	
All partners, European Commission.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	ProCancer-I consortium
Partner in charge of the data analysis	FORTH
Partner in charge of the data storage	FORTH
Related WP(s) and task(s)	WP1
Standards and metadata	
Standards, format, estimated volume of data	Excel format files. Word format files. PDF format files Volume ~1GB
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data.
Access Procedures	None within the project consortium

Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	On project private file repository. Shall be maintained and backed up for a period of 3 years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 13: *DS2.1 & DS2.2 datasets description*

Data identification: *DS2.1 _ Ethical approvals DS2.2_ GDPR documents*	
Generic description:	
The datasets DS2.1 and DS2.2, are related to the legal and ethical issues related to the ProCancer-I project (ethical approvals, informed consents, GDPR documents etc.).	
Origin of data:	
The datasets will be updated during the lifecycle of the project.	
Nature and scale of data:	
The nature of these dataset can be Word and or pdf documents.	
To whom the dataset could be useful:	
All partners related to the data collection and data processing.	
Related scientific publication(s)	
N/A	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	Clinical partners
Partner in charge of the data analysis	Legal and Ethical Committee
Partner in charge of the data storage	Clinical partners
Related WP(s) and task(s)	All WPs
Standards and metadata	
Standards, format, estimated volume of data	Word, PDF
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential (only for members of the Consortium and the Commission Services).
Data sharing, re-use, distribution, publication (How?)	Shall be limited only to be carried out between the Project Consortium members and the European Commission's Services.

Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data are involved. Informed consent will be collected before the proof of concept study starting date.
Access Procedures	None within the project consortium
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The informed consent document will be prepared during work with the clinical project protocol. The consent form will be signed and kept either in a physical format and/or in an electronic format using safe storage platforms on the premises of Clinical centers. Other documents will be stored in the private file repository of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 14: *DS3.1 – DS3-5 datasets description.*

Data identification: ***DS3.1 _ Communication KPIs*** ***DS3.2_ Dissemination materials*** ***DS3.3_ Exploitation plan*** ***DS3.4_ Contact details of linked initiatives*** ***DS3.5_ IPR***
Generic description:
The datasets DS3.1, DS3.2, DS3.3, DS3.4 and DS3.5 include information regarding the dissemination, exploitation and IPR and Innovation management activities of the ProCancer-I consortium.
Origin of data:
The datasets will be updated during the lifecycle of the project, while a final consolidate version of them will be included in the deliverables of WP3.
Nature and scale of data:
The nature of these dataset can be Excel, Word, pdf documents, while the content of the dissemination materials can be web pages, brochure, flyers, PowerPoint presentations, papers in journal and conferences, videos, images etc.
To whom the dataset could be useful:
All partners
Related scientific publication(s)
Journal and conferences publications that will be made during the lifecycle of the project.
Indicative existing similar data sets (including possibilities for integration and reuse):

N/A	
Partners activities and responsibilities	
Partner owner of the data	ProCancer-I consortium partners
Partner in charge of the data analysis	Dissemination and Exploitation Manager
Partner in charge of the data storage	ProCancer-I consortium partners
Related WP(s) and task(s)	All tasks
Standards and metadata	
Standards, format, estimated volume of data	Excel, Word, PowerPoint, PDF Image formats (*.tiff, *.png, *.jpeg etc.) Video formats
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Public for the dissemination materials. Exploitation plan and IPR that will be confidential to the consortium partners and the Commission's Services
Data sharing, re-use, distribution, publication (How?)	The dissemination materials can be shared, re-used and distributed following copyright agreements. Exploitation plan and IPR that shall be limited only to be carried out between the Project Consortium members and the Commission's Services.
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	No personal data are involved.
Access Procedures	For all the public datasets non access procedures are applied.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	Shall be maintained and backed up for a period of 3 years following the end of the project.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

Table 15: *DS5.1 Clinical data description*

Data identification: *DS5.1_Clinical data*
Generic description:

The datasets DS5.1 include patient data (e.g. demographics) medical data with a complete clinical history, disease-related data (PSA, Gleason, etc.) and treatment-related data (treatment type, metastasis, toxicity).	
Origin of data:	
The data will be collected locally at the clinical site of each participating clinical institution.	
Nature and scale of data:	
The nature of these dataset can be Excel or Word documents.	
To whom the dataset could be useful:	
All partners	
Related scientific publication(s)	
Journal and conferences publications that will be made during the lifecycle of the project.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	The clinical partner who provides the data
Partner in charge of the data analysis	Partners responsible to develop the AI models
Partner in charge of the data storage	Clinical and technical partners
Related WP(s) and task(s)	WP5, WP7
Standards and metadata	
Standards, format, estimated volume of data	Excel, Word
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	Sensitive data will be fully anonimized before being distributed to Consurtium members or uploaded to the ProstateNET platform
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	Personal data are involved. Informed consent has be collected from all the participants in the study.
Access Procedures	The access procedures of the data stored at the ProstateNET platform will be described in the next version of ProCancer-I DMP, upon the consotrtium agreement and the DoA requirements.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The anonymazied clinical data that will accompany the DICOM images will be

	uploaded and stored to the ProstateNET platform. The infrastructure and the storage duration will be described in detail in the next Procancer-I DMP version.
Indicative associated costs for data archiving and preservation	N/A
Indicative plan for covering the above costs	N/A

__Table 16:__ DS5.2 Imaging data description

Data identification: *DS5.2_Imaging data*	
Generic description:	
The datasets DS5.2 include multiparametric MRI data from the patients involved in the study.	
Origin of data:	
The data will be collected locally at the clinical site of each participating clinical institution.	
Nature and scale of data:	
The nature of these dataset can be in DICOM format.	
To whom the dataset could be useful:	
All partners working on WP4, WP5, WP6 and WP7.	
Related scientific publication(s)	
Journal and conferences publications that will be made during the lifecycle of the project.	
Indicative existing similar data sets (including possibilities for integration and reuse):	
N/A	
Partners activities and responsibilities	
Partner owner of the data	The clinical partner where the data have been generated.
Partner in charge of the data analysis	Partners responsible to develop the AI models
Partner in charge of the data storage	Clinical and technical partners
Related WP(s) and task(s)	WP4,WP5,WP6, WP7
Standards and metadata	
Standards, format, estimated volume of data	DICOM, Nifti
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	Confidential
Data sharing, re-use, distribution, publication (How?)	Sensitive data will be fully anonimized before being distributed to Consurtium members or uploaded to the ProstateNET platform.
Personal data protection: are they personal data?	Personal data are involved.

If so, have you gained (written) consent from data subjects to collect this information?	Informed consent has be collected from all the participants in the study.
Access Procedures	The access procedures of the data stored at the ProstateNET platform will be described in the next version of ProCancer-I DMP, upon the consotrtium agreement and the DoA requirements.
Embargo periods (if any)	None
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	The anonymazied imaging data will be uploaded and stored to the ProstateNET platform. The infrastructure and the storage duration will be described in detail in the next Procancer-I DMP version.
Indicative associated costs for data archiving and preservation	The actual costs will be defined in the next version of theDMP.
Indicative plan for covering the above costs	N/A

The exact description of the datasets that will be produced from WP4, WP6, WP7 and WP8 will be clarified and described in the next version of the DMP of ProCancer-I.

2.4 Methodology for data collection

Table 1 presents the methodology that is followed for the collection of each one the See Far datasets.

***Table 17**: Datasets collection methodology.*

Datasets of WP1		The datasets of WP1 contain information related to the project management and coordination.
DS1.1	Partners contact list	The data were collected during the initiation phase of the consortium, recorded in an excel file stored in the project's private repository. The partners contact list is updated whenever is necessary
DS1.2	Financial statements	The financial statements information are collected in M6, M12, M18, M24 and M36 following the template provided by the coordinator. They are stored in the project's private repository and in addition, they are included in the interim periodic reported deliverables.
DS1.3	Risk log	The Risk Log for the project is prepared using PM^2 Risk Log template. The PM^2 Risk Log is updated after each

		Project Core-Team Meetings, by the Risk Manager (RM). It is stored in the projects' private repository.
DS1.4	Managerial documents	Managerial documents include the interim periodic progress report documents, the final report of the project, the deliverables, the milestones, the minutes of the conference call and physical meetings, the WP monthly reports etc. The documents are collected as it is planned in the DoA (as far as it concerns the reports, deliverables and milestones), the minutes are collected one week after the end of the conference call and physical meetings, while WP progress reports are prepared by the WP leader and are collected by the coordinator at the end of each month. The managerial documents are stored in the project's private repository and they will be available in the site of the project, in case the dissemination level of the documents allow it.
Datasets of WP2		The datasets of WP2 contain information related to the ethical and legal requirements.
DS2.1	Ethical approvals	The documents concerning ethical approvals will be collected before the starting data of the proof of concept study. A description of the documents as well as the original documents will be presented in the Legal and Ethical Considerations deliverables.
DS2.2	GDPR documents	The documents regarding the application of the GDPR regulation will be prepared, signed by the corresponding partners and collected before the initiation of data collection and processing phase
Datasets of WP3		The datasets of WP3 contain information related to the project dissemination and exploitation.
DS3.1	Communication KPIs	The list of communication KPIs will be completed by the dissemination manager in close cooperation with the partners in M2, M6, M18, M30 and M36 of the project along with the update on the ProCancer-I communication tools.
DS3.2	Dissemination materials	Through the whole lifecycle of the project a list of the dissemination activities along with the content of the dissemination materials is collected and stored in the private document repository of the project while it is uploaded in the website and the social media of the

		project. Additionally, they will be presented in the three versions of the ProCancer-I dissemination plan and activities deliverables (submitted on M12 and M24 and will be submitted on M36).
DS3.3	Exploitation plan	The versions of the exploitation and sustainability plan are created at the end of each year of the project lifecycle.
DS3.4	Contact details of linked initiatives	A document including the information of the linked initiatives has been created and updated in regular basis. A first estimation of the linked initiatives where provided in the DoA.
DS3.5	IPR	IPR Management & Innovation Management documents will be available on M12, M24, M36 of the project.
Datasets of WP4		The datasets of WP4 contain information related to the security, privacy, transparency and sharing of the data repositories and the AI models provided by the ProCancer-I platform.
DS4.1	Data representation	A list of the main ontologies proposed for the radiomics domain and models for data representation that may be used in the ProCancer-I project, has beed drafted. A detail description is provided in the Deliverable 4.2.
DS4.2	Standards for safety and privacy	A list of standards and requirements will be provided after the design of the platform architecture.
Datasets of WP5		The datasets of WP5 contain information related to the data that will be collected for the development of the AI models
DS5.1	Clinical data	Clinical data are mined at the local PACS system of each clinical parner. After curation and anonymization, clinical data will accompany each imaging study that will be used for generating the ProstateNET dataset.
DS5.2	Imaging data	DICOM imaging series and annotation masks are extracted by authorized local users from the various hospital information systems (PACS) where they were generated. Then, as the technological infrastructure would be available for data curation, annotation and upload, the ProstateNET dataset will be populated.
Datasets of WP6		The datasets of WP6 contain information related to ethical, trustworthy and FAIR aspects of AI models.
DS6.1	Performance monitoring	Initially, models' performance will be recorded during model training (WP5 & WP6) before uploading them to the ProstateNET platform.

DS6.2	Radiomics and AI FAIRification	Data FAIRness will be evaluated and recorded through the whole lifecycle of the project for all generated data than need to comply with the FAIR principles.
Datasets of WP7		The datasets of WP7 contain information related to the performance the AI models in the context of 8 clinical scenarios.
DS7.1	Error tracking	A list of errors and mispecifications will be created after aploading the modles to the platform and will be updated when necessary.
DS7.2	Safety and effectiveness	Models' performance will be recorded during model validation (WP7) and will be updated regularly after uploading them to the ProstateNET platform.
Datasets of WP8		The datasets of WP8 contain information related to the sustainability planning and business opportunities.
DS8.1	Sustainability & Business plan	The documents concerning sustainability plan and business opportunities will be collected along the course of the project. The documents will be stored in the project's private repository and they will be available in the site of the project, in case the dissemination level of the documents allow it.

2.5 Data for AI model development

For the development and validation of the AI models mpMRI, and clinical data, retrospectively and prospectively, from more than 17.000 PCa patients (11.000 retrospective and 6.000 prospective mpMRI cases), will be exploited, counting more than 1.5 million prostate image representations. The collection of mpMRI scans, include a high-resolution T2-weighted imaging and at least two physiology-based MRI techniques (diffusion-weighted (DW) and dynamic contrast-enhanced (DCE) imaging) for every participant. Additionally, each mpMRI case will be accomplanied by non-imaging data (clinical, toxicity, with the aim to generate a specific disease profile that will be linked to specific AI outcomes and will determine the individualized patient response with respect to 9 clinical questions.

The data that will be collected for training and validation of the AI models, with respect to each clinical use case, are reported in Table 18.

Table 18: Non-imaging data collected for the implementation of the AI models.

Use Case	Description	Collected data
UC1	Detection of PCa with high accuracy both in peripheral and transitional zone	Clinical (age, DRE, clinical history, etc), PSA, PSAD, Gleason Group.
UC2	Characterization of cancer according to its biological aggressiveness	Clinical (age, DRE, clinical history, etc), PSA, PSAD, Gleason Group.
UC3	Early identification of patients with metastatic prostate cancer	Clinical (age, DRE, clinical history, etc), PSA, Gleason group, status of resection margins, presence of extra prostatic invasion, N status, post-RP PSA.
UC4	Radiologic – Histopathologic correlation	Clinical (age, DRE, clinical history, etc), PSA, Gleason group.
UC5	Prediction of the risk of local disease recurrence after radical prostatectomy	Clinical (after treatment), PSA after treatment, Gleason group, status of resection margins, presence of extraprostatic invasion, nodal status.
UC6	Prediction of risk of disease recurrence after radiation therapy.	Clinical (after treatment), PSA after treatment.
UC7	Prediction of post radical prostatectomy and/or radiation-induced urinary toxicity:	PSA after treatment, toxicity data after radiation treatment, QoL assessment after treatment, dose/RT plan, no. of RT sessions.
UC8	Patient stratification for enrollment in Active Surveillance programs	Clinical, Gleason Grade at baseline, eventual other pathological variables.
UC9	Prediction of the best treatment option with lowest side effects/toxicity	Clinical (before and after treatment), PSA after treatment, Gleason group, status of resection margins, presence of extraprostatic invasion, nodal status, toxicity data after radiation treatment, QoL assessment after treatment, dose/RT plan, no. of RT sessions.

2.6 Storage and Back up

In ProCancer-I project the aforementioned datasets will be stored either to the project private repository and to the prostateNET repository which is part of the ProCancer-I secure cloud platform that also integrates several valuable services and beyond the state of the art algorithms related to the clinical scenarios described in the description of action of the project. The project private repository is based on the CBMLBox[2] which is a cloud based repository based on the Nextcloud repository technology.

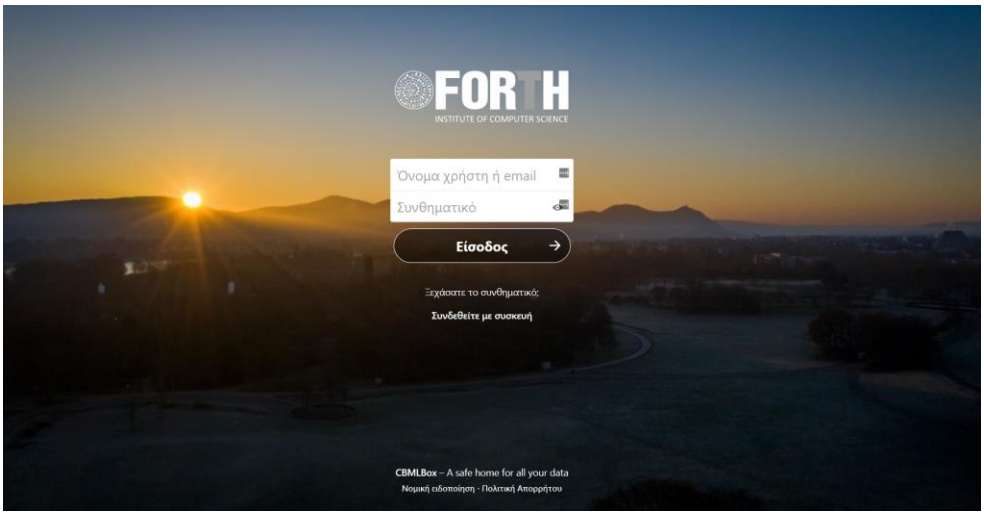

Figure 1: *Login landing page of the ProCancer-I private repository*

Moreover, the prostateNET repository is built on top of the centralised ProCancer-I platform infrastructure. All data will be residing in a data lake including both the imaging data and the clinical datasets and their annotations. Data assets will be shown on a semantically harmonized, searchable and browsable metadata catalogue based on open standards including FAIR ones (findability, accessibility, interoperability and reusability). Rich metadata, semantic annotation, unique identifiers and versioning functions will be included along with secure and privacy-preserving metadata publication and search. The consortium will build on the experience in metadata repository, distributed ledgers for metadata registration and tracking and discovery indexes. In those metadata repositories, ontologies and other related standards/terminologies will be used, semantically uplifting and homogenizing the available data. All those data will be accessed by the machine learning, AI and analytical tools through APIs offered by the data storage layer.

[2] https://cbmlbox.ics.forth.gr/index.php/login

2.7 Data security

Security is of high importance for the ProCancer-I consortium which is a multifaceted quality attribute that affects the functionality, user experience, availability, and data protection. For both the private and ProstateNET repositories the following aspects are addressed:

- **Authentication**: provide information about the person or the system interacting with the platform
- **Authorization**: restrict access control based on the users' identity and their access rights on the data and compute resources
- **Audit**: logging and monitor of users' and systems' behaviour throughout the platform
- **Confidentiality**: all interactions are encrypted and protected from unauthorized access and eavesdropping

Specifically for the ProCancer-I cloud platform, all the information in transit will be performed over secure channels. The use of Transport Layer Security (TLS) with strong ciphers is the established best practice for securing network communication. Additionally TLS provides integrity and authenticity of the interacting peers. In addition, the TLS-secured network communication channels and HTTPS will be used both internally (among the platform's components) and externally (when the system is accessed by its users or other systems). The identification and authenticity of the interacting ProCAncer-I components will be verified with digital certificates signed by either using well known and trusted Certificate Authorities (CA) or an internal CA of the platform in order to simplify deployment. The latter can facilitate the testing of the components and it's certainly easier, at the cost of supporting only the internal communication. Of use of strong private keys (2018-bit RSA or 256-bit ECDSA), recent versions of TLS (TLS 1.2 and 1.3), and a short list of strong ciphers that offer at least 128-bit encryption will be utilized.

Regarding the private repository based on the Nextcloud technology integrates logging and intrusion detection and works with existing authentication mechanisms like SAML, Kerberos and LDAP. Administrators can set permissions on sharing and access to files using groups. The CBMLBoc employs standard TLS to encrypt data in transfer and offers Server Side Encryption on the local storage

2.8 Patient's data anonymization

For the ProCancer-I, the anonymisation is a mandatory process before any medical image and the accompanied clinical data are distributed oustside of the clinical site where thery were generated. In ProCAncer-I, DICOM image anonymization will be performed in order to securily and definitevely remove any personal information included in the DICOM tags or even in the pixel data (pixel black-out support). The DICOM anonymization tool will be used to anonymize DICOM images at the clinical premises before they are uploaded to the ProCAncer-I platform. The DICOM anonymization tool implements a whitelist solution in order to cope with the personal health information that might be included in the DICOM tags. The whitelist has been defined by ProCAncer-I's partners involved in the anonymization task, and its purpose is filtering the DICOM

tags preserving only the ones included in the whitelist (either with their original values, either modified, depending on each respective tag and based on a predefined set of rules) while removing the others. The DICOM anonymization tool is a standalone application that has no other automatic interaction with other tools/components, as the anonymization process is a manual procedure performed by the clinician. Besides tools for the anonymization of the DICOM data, there will also be available tools for the anonymization of the clinical data that will accompany the DICOM series. Upon anonymization, the data upload tool will be installed at the clinical sites premises to upload the data to the staging area of the ProCancer-I platform. The sequence diagram of the aforementioned process is shown in Figure 2.

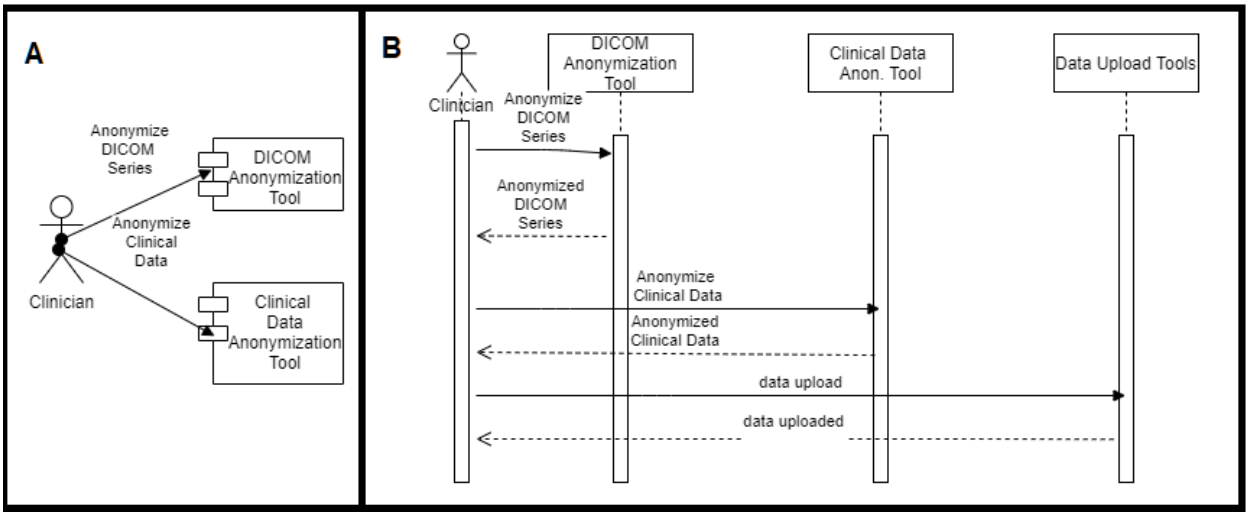

***Figure 2**: The two individual components of the data anonymization tools (A) and the sequence diagram for anonymization and uploading data to the platform (B).*

3 Alignment to the Findable, Accessible, Interoperable, Re-usable (FAIR) data principles

ProCAncer-I aims to provide the global reference data resource, data specifications and format and application platform in support of the next generation PCa AI research. All consortium partners have been committed to ensure that the data produced, collected and processed, align with the FAIR Principles definition[3], thus be findable, accessible, interoperable and reusable. Through the life cycle of the ProCancer-I project, the FAIR principles will be followed as far as possible, for both the data available in the infrastructure and the AI models, while ensuring compliance with national and European ethic-legal framework.

This section is based on the guidelines for effective data management in in the course of a Horizon 2020 project, provided by the European Commission[4]. The FAIR component of the current DMP version comprises points to clarify, which will be addressed during the course of the project.

The proposed FAIRification workflow, tailored to the specific needs and requirements posed by the use of health data, is shown in Figure 3. It is based on the GO FAIR[5] initiative, extended by FAIR4Health project[6], to meet specifically the challenges of health data for being FAIR.

3.1 Making data Findable, including provisions for metadata

The first step in the FAIRification process is to easily find them inside the large data pools. Thus, both metadata and data should be easily recognisable by both humans and machines.

All the deliverables will be listed on the ProCancer-I website (www.procancer-i.eu), and the ways by which ProCancer-I output can be accessed will be communicated via social media and other suitable channels to increase visibility of ProCancer-I work. For public deliverables, a link will be available between the ProCancer-I website and the appropriate open repositories where the data is submitted.

[3] Wilkinson M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 doi: 10.1038/sdata.2016.18 (2016).
[4] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-accessdata-management/data-management_en.htm
[5] https://www.go-fair.org/go-fairinitiative/
[6] Sinaci A.A. et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020 Jun;59(S 01):e21-e32. doi: 10.1055/s-0040-1713684

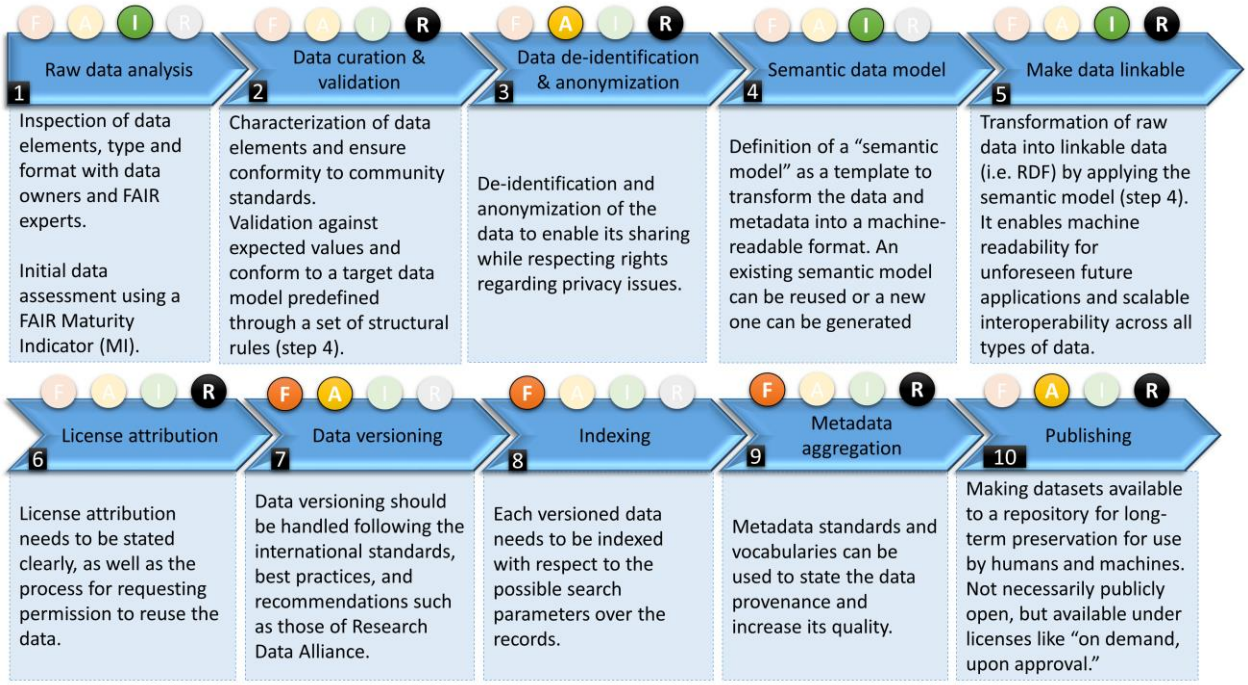

Figure 3: *The FAIRification workflow for the ProCancer-I data adjusted from Sinaci et al.[7]. Each step is characterized based on the FAIR requirements it is addressing ((F)indability, (A)ccessibility, (I)nteroperability, and (R)eusability)*

3.1.1 F1. Assign globally unique and persistent identifiers to data and metadata

Interpretation: Principle F1 states that digital resources, i.e., data and metadata, must be assigned a globally unique and persistent identifier in order to be found by computers. Globally unique and persistent identifiers remove ambiguity in the meaning of published data by assigning a unique identifier to every element of metadata.

Implementation: For openly available data produced by the project, such as scientific publications, a Digital Object Identifier (DOI) will be issued directly by Zenodo once they are uploaded to the Zenodo repository[8]. The assignment and management of the persistent and globally unique identifiers (produced by the ProCancer-I platform, the ProstateNET), addressed in this first version of the ProCancer-I DMP, will be updated through the whole lifecycle of the project.

3.1.2 F2. Describe the project data with rich metadata

Interpretation: Rich metadata, including descriptive information about the context, quality and characteristics of the data allows finding data based on the information provided by their metadata, even without the need of the identifier. To enable both global and local search engines

[7] Sinaci AA et al. From Raw Data to FAIR Data: The FAIRification Workflow for Health Research. Methods Inf Med. 2020 Jun;59(S 01):e21-e32. doi: 10.1055/s-0040-1713684

[8] http://about.zenodo.org/

to locate a resource, generic and domain-specific descriptors should be provided.

Implementation: Concerning imaging data, metadata will be (1) automatically attached to all medical images (confined within the DICOM format) and (2) collected via a graphical user interface for the researcher. These metadata, after anonymization and de-identification, will be uploaded to the ProstateNET repository.

Wherever possible, curated datasets will be published on Zenodo which makes them openly accessible and discoverable. In addition, data will also be made available via the ProCancer-I website accompanied with the respective metadata. The data that will be uploaded to Zenodo will inherit the metadata description of Zenodo, which is compliant with DataCite's Metadata Schema minimum recommended terms, with a few additional enrichments. The DataCite Metadata Schema for Publication and Citation of Research Data allow data to be understood and reused by other members of the research group and add contextual value to the datasets for future publishing and data sharing. Text files metadata will be automatically generated using the DataCite Metadata Generator after filing the form requesting intrinsic metadata. All published data sets will receive a DOI that will be referred to in any scientific publication that made use of this data set. A list of metadata elements required for data citation are given in Figure 2.

***Table 19**: Citation and Discovery metadata for data repositories across common standards. (Source: M.Fenner et al., 2019. A data citation roadmap for scholarly data repositories. Nature Scientific Data)*

Citation Metadata	**Dublin Core**	**Schema.org**	**Datacite**	**DATS**
Dataset Identifier	Identifier	@id	Identifier	identifier
Title	Title	Name	Title	Title
Creator	Creator	Author	Creator	creator
Data repository or archive	Publisher	Publisher	Publisher	Publisher
Publication Date	Date	datePublished	publicationYear	Date
Version	N/A	Version	Version	Version
Type	Type	Type	resourceTypeGeneral	type
Discovery Metadata	**Dublin Core**	**Schema.org**	**Datacite**	**DATS**

Description	description	description	description	datatype dimention Material …
Keywords	Subject	Keywords	Subject	keywords
License	Lisence	License	Rights	licence
Related Dataset	isPartOf isVersionOf references	isPartOf citation	relatedIdentifier	isPartOf
Related Publication	bibliographicCitation	Citation	relatedIdentifier	publication

3.1.3 F3. Clearly and explicitly include in the metadata the identifier of the data they describe

Interpretation: Principle F3 states that any description of a digital resource must contain the identifier of that resource being described. This is especially important where the resource and its metadata are stored independently, but persistently linked, which is generally considered good practice in FAIR.

Implementation: The association between a metadata file and the dataset will be made explicit by mentioning dataset's globally unique and persistent identifier in the metadata. Where applicable for specific ProCancer-I datasets the metadata creation will be based on widely used standards. To guarantee that the connection is annotated in a formal manner, the FAIRifier tool[9] may be used. Alternatively, the FAIR Data Point, which is based on the Data Catalogue model (DCAT)[10], provides not only unique identifiers for potentially multiple layers of metadata, but also a single, predictable, and searchable path through these layers of descriptors, down to the data object itself. The most prominent metadata catalogues considered by the ProCancer-I project for collecting, managing, analyzing, visualizing and sharing data are Egeria[11], MOLGENIS[12], CKAN[13], InvenioRDM[14] and BBMRI-ERIC[15]. Among these metadata catalogs, the MOLGENIS metadata catalogue seems to be the most appropriate for the ProCancer-I project given the already high adoption of the community, its modular and extensible design and the already available modules.

[9] https://github.com/FAIRDataTeam/OpenRefine-metadata-extension/
[10] https://www.w3.org/TR/vocab-dcat/
[11] https://egeria.odpi.org/
[12] https://www.molgenis.org/
[13] https://ckan.org/features/
[14] https://inveniosoftware.org/products/rdm/
[15] https://www.bbmri-eric.eu/bbmri-eric/common-service-it/

3.1.4 F4. Register or index the data in a searchable resource

Interpretation: Principle F4 states that digital resources must be registered or indexed in a searchable resource. The searchable resource provides the infrastructure by which a metadata record (F1) can be discovered, using either the attributes in that metadata (F2) or the identifier of the data object itself (F3) [21].

Implementation: Metadata of each record uploaded to Zenodo is indexed directly in Zenodo's search engine, immediately after publishing. Metadata of each record is sent to DataCite servers during DOI registration and indexed there[16]. Datasets can also be indexed with metadata in common search indexes, such as Google Dataset Search via schema.org[17]. Keywords will be provided, based on standard terminologies, potentially enabling multilingual search, and search at various levels of detail. In addition, metadata may be shared via FAIRSharing[18] or other relevant findability service providers that deliver both human- and machine-readable access to metadata.

3.2 Making data Accessible

This principle refers to easy access of the data from the user, possibly including authentication and authorization. Since some ProCancer-I data can be confidential they will be restricted in their use. Sensitive and personal data can be made accessible only following the GDPR requirements. As several repositories will be used to store data, the policy on how to grant access to restricted results will be developed over the course of the project and described in future DMPs.

3.2.1 A1. (meta)data are retrievable by their identifier using a standardized communications protocol

Interpretation: Principle A1 states that FAIR data retrieval should be mediated without specialised or proprietary tools or communication methods and that the identifier (F1) follows a globally-accepted schema tied to a standardized, high-level communication protocol. Its purpose is to provide a predictable way to access a resource, regardless of whether unrestricted access to the content of the resource is granted or not. Conditions of compliance are further specified in sub-principles A1.1 and A1.2.

A1.1 the protocol is open, free, and universally implementable

Implementation: The most common example of a compliant standardized access protocol is the Hypertext Transfer Protocol (HTTP[19]). It offers useful features, including the ability to request

[16] http://about.zenodo.org/principles/
[17] https://datasetsearch.research.google.com/
[18] https://fairsharing.org/
[19] https://www.w3.org/Protocols/

metadata in a preferred format, and/or to inquire as to the formats that are available. It is also widely supported by software and common programming languages.

A1.2 the protocol allows for an authentication and authorization procedure, where necessary

Implementation: The software developed during the project will be hosted on a Git-like server. Documentation and a user guide with examples will be published as an online tutorial via the website and will accompany any release.

As specified to the IPR policy of the consortium, the collected data will be shared with the scientific community, when it does not concern sensitive information. Regarding clinical data, for confidentiality reasons, data will be anonymised before being shared even if such sharing takes place only with the consortium. The direct access for some ProCancer-I datasets will be accomplished through a private repository based, for example, on the Dropbox cloud technology. It is a well- known secure and powerful repository with many features in syncing, backup and data file handling. In this case, http(s) or ftp communication protocols will be used to provide access to these data, making them free, open-sourced and globally implementable, facilitating thus data retrieval.

All the others datasets for ProCancer-I project will securely uploaded to the ProstateNET, a centralised platform which will be designed and developed during the project lifecycle. For data generated within ProCancer-I the Data Owner/Data Provider should agree that these are of high level of granularity and metadata description. Prior to the data collection any ethical and legal compliance will be assessed and confirmed. Moreover, for the data collected outside the project (existing – not generated internally) the necessary quality assessment will be performed and anonymization process will be performed if necessary. For highly sensitive data, an email, telephone number, or Skype name of a contact person who can discuss access to the data will be provided. This contact protocol will be clear and explicit in the metadata. Hence, even heavily protected and private data will be FAIR.

3.2.2 A2. metadata are accessible, even when the data are no longer available

Interpretation: In case where the data record is no longer available, there must be a clear and precise way of discovering its historical metadata record. Collected, processed and generated research should be uploaded and preserved in wide acceptable formats so as to ensure long-time accessibility. The selection of the adequate data format needs to be carefully assessed in order to have a high chance of being usable in the future.

Implementation: For publicly accessible ProCancer-I data through the open repository Zenodo, the consortium will use the provided metadata for individual records as well as record collections which are harvestable using the OAI-PMH protocol and also retrievable through the public REST API. Both are open, free and universal protocols for information retrieval on the web. These data and metadata stored in Zenodo will be retained for the lifetime of the repository which for now is the lifetime of the host laboratory CERN, the next 20 years at least.

3.3 Making data Interoperable

3.3.1 I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

Interpretation: The purpose of principle I1 is to achieve a "common understanding" of digital resources through a globally understood language for machines, with an emphasis on "knowledge" and "knowledge representation".

Implementation: To ensure automatic findability and interoperability of datasets, commonly used controlled vocabularies, ontologies and thesauri will be used to describe the dataset. The most widely-accepted framework to describe and structure (meta)data is, currently, the Resource Description Framework (RDF) extensible knowledge representation model which is the W3C's recommendation for how to represent knowledge on the Web in a machine-accessible format[20] . It provides a common and straightforward underlying model and creates a powerful global virtual knowledge graph.

3.3.2 I2. (meta)data use vocabularies that follow FAIR principles

Interpretation: Principle I2 uses "vocabularies" to refer to the methods that unambiguously represent concepts that exist in a given domain. I2 requires that the vocabulary terms used in the knowledge representation language (principle I1) can be sufficiently distinguished, by a machine, to ensure detection of "false agreements" as well as "false disagreements".

Implementation: In order to make the data interoperable in ProCancer-I, standard open formats will be used for storage. The necessary vocabularies and wide-open standards will be used to design the data schema and store the data. Proprietary software and language-dependent formats will be avoided where possible.

Ontologies defined in the "Web Ontology Language" (OWL) and shared via a publicly accessible registry (e.g. BioPortal for life science ontologies[21]) are examples of formally represented, accessible, mapped, and shared knowledge representations in a broadly applicable language for knowledge representation. In the radiomics domain, the IBSI[22], Radiomics Ontology (RA)[23] and Radiation Oncology Ontology (ROO)[24] are some model representations widely supported and developed by the community that are also compliant with the Findability requirements of FAIR.

Since the project will collect a large number of medical images, we will use that domain-specific standard (DICOM) as a starting point for those data. For other data, such as tabular datasets, CSV,

[20] https://www.w3.org/RDF/
[21] https://bioportal.bioontology.org/
[22] Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., ... & Löck, S. (2020). The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology, 295(2), 328-338.
[23] https://bioportal.bioontology.org/ontologies/RO
[24] https://bioportal.bioontology.org/ontologies/ROO

XML or JavaScript Object Notation (JSON) will be used to format the metadata. In situations where wider standards, such as Dublin Core, are needed, we will provide proper mappings.

3.3.3 I3. (meta)data include qualified references to other (meta)data

Interpretation: A "qualified reference" is a reference to another resource (i.e., referencing that external resource's persistent identifier), in which the nature of the relationship is also clearly specified. This principle therefore relates to the good practice to clearly distinguish between metadata (files/containers) and the resources they describe.

Implementation: The scientific links between the datasets will be described, specifying if one dataset builds on another data set, if additional datasets are needed to complete the data, or if complementary information is stored in a different dataset. Furthermore, all datasets will be properly cited (i.e., including their globally unique and persistent identifiers).

3.4 Making data Reusable

The ultimate goal of FAIRification process is to optimise the reuse of data. To accomplish this, metadata and data should be well-described and appropriately licensed so that they can be replicated and/or exploited in different settings.

3.4.1 R1. (meta)data are richly described with a plurality of accurate and relevant attributes

The focus of R1 is to enable machines and humans to assess if the discovered resource is appropriate for reuse, given a specific task. This reiterates the need for providers to consider not only high-level metadata facets that will assist in generic search (as described by principle F2),but also to consider more detailed metadata that will provide much more "operational" instructions for re-use. The sub-principles R1.1, R1.2 and R1.3 define some critical types of attributes that contribute to R1.

ProCancer-I is expected to generate a substantial volume of novel data and knowledge through its lifecycle. For datasets and the scientific articles freely available in the Zenodo repository, the re-use principle is inherently supported since: (1)Each record contains a minimum of DataCite's mandatory terms, (2) License is one of the mandatory terms in Zenodo's metadata and is referring to an Open Definition license, (3) Data downloaded by the users is subject to the license specified in the metadata by the uploaded, and (4) Zenodo is not a domain-specific repository

3.4.2 R1.1. (meta)data are released with a clear and accessible data usage license

Interpretation: Digital resources and their metadata must always, without exception, include a license that describes under which conditions the resource can be used, even if that is "unconditional". In order to facilitate reuse, the license chosen should be as open as possible.

Implementation: ProCancer-I team recognizes the importance of software licensing from the outset of the project. Therefore, (meta)data will be released with a clear and accessible data usage license, like MIT (for software codes) or Creative Commons (for datasets) since these are the best option between unrestricted access and the promotion of a fair community practice that acknowledges the provenance of data. In the cases where specific service information cannot be publicly shared, the reasons will be mentioned in their metadata descriptions (e.g. ethical, personal data, intellectual property, commercial, privacy-related, security-related). Nevertheless, given the uncertainty of the potential value of the developed tools in the future, the exact licenses to be used in case-by-case situation will be refined through the course of the project. At this point, we are exploring the spectrum of licenses as a preliminary step based on the information found at https://choosealicense.com/. Once we will have a clearer idea of what type of data will be produced during the project (e.g. open sources tools, mix of proprietary and software tools, etc.), the DMP will be updated with relevant information. Also, the embargo time, how and when third parties will gain acess to selected datasets and the time duration for data re-use are elements to be specified.

3.4.3 R1.2. (meta)data are associated with detailed provenance

Interpretation: Detailed provenance includes facets such as how the resource was generated, by whom, under what conditions, using what starting data or source-resource, using what funding/resources, who owns the data, who should be given credit, and any filters or cleansing processes that have been applied post-generation.

Implementation: Provenance descriptions can be implemented following community specific templates according to the PROV-Template[25] approach. These templates allow to predefine the structure of the intended collection of provenance information using variables which are later instantiated with appropriate data extracted from existing process output. Another good example of R1.2 metadata schemas are those composed as CEDAR[26] templates, which offer community-defined FAIR metadata elements and FAIR metrics for building machine-actionable metadata profiles.

3.4.4 R1.3. (meta)data meet domain-relevant community standards

Interpretation: When formal community standards or best practices for data archiving and sharing exist, they will be followed. Minimal Information Standards are describing most often the minimal set of metadata items required to assess the quality of the data acquisition and processing and to facilitate reproducibility.

Implementation: Metadata registries, offering a list of community standards, will be used to choose domain-specific standards taking into full consideration the relevant inter-domain

[25] https://provenance.ecs.soton.ac.uk/prov-template/
[26] https://more.metadatacenter.org/tools-training/outreach/cedar-template-model

interoperability requirements. FAIRsharing[27], Identifiers.org[28] and Bioportal[29] are examples of metadata registry for metadata standards, including file formats, ontologies, identifier schemas, as well as for maturity Indicators.If there is a lack of metadata standards, the consortium will reuse existing controlled vocabularies for providing metadata to resources as far as possible. Controlled vocabularies for reused can be found on Joinup[30] and Linked Open Vocabularies[31] platforms. If there is no suitable authoritative reusable vocabulary for describing data, conventions will be used for describing the vocabulary: RDF Schema (RDFS) and/or Web Ontology Language (OWL).

3.5 Implementation of data FAIRification

A growing tool kit for FAIR is available to facilitate the construction of FAIR (meta)data (some under development). The implementation of the fair principles require a combination of expert training (data stewards), content (FAIRsharing[32]) technology (Data Stewardship (DS) Wizard[33]; Castor EDC[34]. CEDAR[35], FAIR Data Points[36], FAIR Metrics Evaluators), and trusted third-party certification organisations (GO FAIR Foundation).

Both CEDAR and Castor EDC electronic case-report form systems can be used to capture FAIR machine-actionable data and metadata throughout the project. Because CEDAR templates can be created to satisfy or validate many of the FAIR principles as users enter their metadata, the metadata-entry process will provide immediate feedback that will help make the metadata more FAIR, including indications of what steps must be be taken to improve the metadata. Castor EDC Forms and Form Templates can be annotated with metadata, including automatically consuming metadata produced within CEDAR and the DS Wizard.

A detailed list of tools available for facilitating the implementation of data FAIRification are provided in the Apendix II.

3.6 Evaluation of data FAIRness

A final step in the post-FAIRification phase is to assess the FAIRness of the data. This process may include: 1) an evaluation to check whether the original objectives have been achieved (if not, some of the steps in the workflow may need to be revisited), and 2) checking the FAIR status of

[27] https://fairsharing.org/standards/ https://doi.org/10.1038/s41587-019-0080-8
[28] www.identifiers.org
[29] www.bioportal.org
[30] http://joinup.ec.europa.eu
[31] http://lov.okfn.org
[32] https://fairsharing.org/
[33] https://ds-wizard.org
[34] https://www.castoredc.com/for-researchers/
[35] https://metadatacenter.org/
[36] https://www.dtls.nl/fair-data/find-fair-data-tools/

the data and metadata using FAIRness assessment tooling. Thus, tools developed for conducting FAIRness evaluations can be based on either discrete/ open-answer evaluation questionnaires or semi-automated evaluation models[37].

Communities have already published documents that can guide implementation choices. Some examples are "the FAIR metrics"[38] and the follow-up Maturity Indicators[39] , "Top 10 FAIR Data & Software Things"[40] and the FAIR Convergence Matrix[41]. Self-assessment models for measuring the maturity level of a dataset have also been developed such as the RDA FAIR Data Maturity Model[42]. On top of that, some evaluator services not only run the FAIR metrics evaluations, but also are intended to deliver a certified report regarding compliance with the FAIR Principles and the resulting level of FAIRness[43].

[37] R. de Miranda Azevedo & M. Dumontier Considerations for the Conduction and Interpretation of FAIRness Evaluations. Data Intelligence (2020)
[38] M.D. Wilkinson et al. Comment: A design framework and exemplar metrics for FAIRness. Scientific Data (2018).
[39] M.D. Wilkinson. Evaluating FAIR maturity through a scalable, automated, community-governed framework. bioRxiv (2019)
[40] C. Erdmann et al. Top 10 FAIR data & software things. Zenodo (2019)
[41] H.P. Sustkova et al. FAIR convergence matrix: Optimizing the reuse of existing FAIR-related resources. Data Intelligence (2020); https://www.go-fair.org/today/fair-matrix/
[42] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[43] https://www.gofairfoundation.org/certification/

4 General Data Protection Regulation (GDPR)

4.1 General

EU citizens are granted the rights of privacy and data protection by the Charter of Fundamental rights of the EU. In particular, Art. 7 states that "everyone has the right respect for private and family life, home and communications", whereas Art. 8 regulates that "everyone has the right to the protection of personal data concerning him or her" and that processing of such data must be "on the basis of the consent of the person concerned or some other legitimate basis laid down by law." These rights are developed in detail by the General Data Protection Regulation (GDPR), Regulation 2016/679/EC, which is in force in every Member State since 25th of May of 2018[44].

The GDPR applies only to the processing of personal data. Since the EU data protection legislation only deals with the processing of personal data, the distinction of personal and non-personal data (which includes anonymous data) is crucial for all activities of the project. Article 4(1) GDPR defines personal data as:

any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

The use of anonymized, synthetic, dummy, fake or any other data that does not qualify as personal data, as it is does not relate ton an identified or identifiable natural person and therefore the duties and potential liabilities associated with the GDPR are removed in the processing of anonymised data. Furthermore, this processing eliminates any possible risks to the rights and freedoms of data subjects as it is no longer possible to associate the data with a data subject. Therefore, the use of non-personal data is strongly advised whenever this is possible and considered functional for the purposes of the processing.

In the ProCAncer-I project, the Partners decided to exclusively work with anonymized data. Therefore, the GDPR will not apply to this data processing. The data will be anonymized by the clinical partner who originally collected the personal data. However, the anonymization process, where personal data is transformed from personal to non-personal data is considered data processing; therefore, this step lies within the scope of the GDPR.

In the following sections, the legal basis for the anonymization will be briefly described and the consortium will make a distinction for retrospective and prospective data.

[44] https://eur-lex.europa.eu/content/news/general-data-protection-regulation-GDPR-applies-from-25-May-2018.html

4.2 Legal requirements of anonymization

The requirements for anonymization are not definitively outlined in the GDPR, which is why the possibility for incomplete anonymization was pointed out as one of the main risks in the DPIA.[45] The main legal conditions for a GDPR-compliant anonymization are described below.

As mentioned above, the GDPR is only applicable to the processing of personal data as set out in Article 1(1). Recital 26 GDPR explicitly states that: "*The principles of data protection should therefore not apply to anonymous information, namely information 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. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes*."

The Article 29 Data Protection Working Party ('WP29') describes anonymization as a "*technique applied to personal data in order to achieve irreversible de-identification*".[46] However, the process of anonymization is often confused with processes of pseudonymization. Personal data which has undergone pseudonymization and which could be attributed to a natural person with the use of additional information, cannot be described as anonymized for GDPR purposes[47] as it allows for the re-identification of the data subject. Many of the pseudonymization techniques commonly used in health research, e.g. key coding of data, may not be sufficient to take the processing outside of the scope of the GDPR requirements.[47] There is a degree of uncertainty as regards the borderline between pseudonymizaiton and anonymization for GDPR purposes.[47] The primary distinction between pseudonymizaiton and anonymization is that the former allows for the re-identification of the data subject, whereas the latter process does not.

In assessing the possibility and probability of the re-identification of the data subject, Recital 26 of the GDPR requires an objective assessment of the measures which are likely enable re-identification, such as the costs of identification, the time required for identification and the available technology and foreseeable technological developments.

To fulfill the above-mentioned legal requirements and ensure anonymization, the Partners shall not store any information which allows for the identification, whether directly or indirectly of the concerned data subject. The Partners shall utilize the anonymization tools, which are agreed upon within the Consortium. Furthermore, the Consortium shall review regularly whether their anonymization methods reflect the current state of the art.

[45] Data Processing Impact Assessment, D2.3.
[46] *Donnelly/McDonagh*, Health Research, Consent and the GDPR Exemption, European Journal of Health Law 26 (2019) 100; Opinion, 05/2014 on Anonymisation Techniques, adopted 10 April 2014, 0829/14/EN, 7.
[47] *Donnelly/McDonagh*, Health Research, Consent and the GDPR Exemption, EJHL 26, 100.

4.3 Retrospective data

The Consortium concluded that the legal basis for the anonymization process of the clinical data will be Article 9(2)(j) GDPR, also known as the research exemption.[48] The legal, ethical and practical advantages – in comparison to the retrospective consent of the former patients – were concluded in the ProCAncer-I RETRÒ Protocol.[49] A short summary of the main arguments and challenges may be found below.

The first and most straightforward legal basis for the processing of personal data here is the explicit consent of the data subject in accordance with Article 9(2)(a) GDPR. However, as this protocol is dealing with retrospective data, which was mainly collected prior to the entry into force of the GDPR, the obtainment of explicit consent may be problematic from a legal, ethical and practical point of view. Therefore, an alternative legal basis that the Consortium could rely on is the so-called 'research exemption'.

The GDPR's research exemption is set out in Article 9(2)(j) GDPR. This provision states that the prohibition on the processing of special categories of personal data (which includes health data) does not apply where the processing is necessary for scientific research purposes. In addition, processing has to be "*in accordance with Article 89(1) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject*".[48] If these requirements are fulfilled, this provision serves as an alternative to the explicit consent requirement, as set out in Article 9(2)(a) GDPR. Article 9(2)(j) must be read together with the standards for lawful processing under Article 6(1).[50] The Consortium is convinced that there are substantive legal arguments that the conditions set out in Article 9(2)(j) and 6(1) GDPR are fulfilled.[49] Nevertheless, it has to be kept in mind that processing based on Article 9(2)(j), must acknowledge the safeguards in Article 89(1) and have a basis in Member State law. As such, there may be deviations in national law. Consequently, Partners processing personal data based on Article 9(2)(j) GDPR are required to check the above-mentioned requirements with their DPOs in order to comply with their national legislation.

Given the retrospective nature of the study and the large number of datasets, obtaining consent from all former patients would be extremely time-consuming, resource intense, and detract from the research itself, thereby negatively affecting ProCAncer-I project in terms of timing and research value and output.

From an ethical point of view - considering the retrospective nature of the study – partners have agreed that the requirement to obtain informed (ethical) consent can be waived. It is important to point out that the informed (ethical) consent is different from the consent to the processing of personal data pursuant to the GDPR. Informed (ethical) consent can be described as the process of informing the potential research participants about the key elements of a research

[48] Article 9(2)(j) GDPR.
[49] ProCancer-I retrospective protocol.
[50] *Donnelly/McDonagh*, Health Research, Consent and the GDPR Exemption, EJHL 26, 112.

study and what their participation will involve, thereby allowing them the autonomy to freely decide on their participation. The informed consent process is one of the central components of the ethical conduct of research with human subjects, however waiving of the requirement of informed consent can be justified by several factors:

- the number of subjects that will be collected is very large, therefore, the effort, time and resources necessary to contact each individual would be unreasonable and would negatively affect the ProCAncer-I research;
- the study deals only with retrospective data that is already available at the clinical Partner's institution and it will be completely anonymized, so that the enrolled subjects cannot be re-identified in any way. The study does not require new analyses or acquisitions of follow up data;
- given the retrospective nature of the study, all patients have already completed their treatments. It is considered appropriate to protect patients from being aware about the study, since it may cause psychological distress, associated with the recollection of their illness status, pain and treatments;
- given the advanced age of some groups of patients, death may have occurred in some cases. Therefore, contacting the family could also cause distress in remembering the event.

Before initiating the trial, the clinical Partners should have submitted the protocol favorable opinion from the EC/CA for the trial conduction. All the correspondence with the EC/CA should be retained in the Investigator File. Before implementing any protocol amendment, the EC must be notified.

The trial will be performed in accordance with International Conference on Harmonization Good Clinical Practice guidelines, the Declaration of Helsinki and applicable local regulatory requirements and laws.

4.4 Prospective Data

For the collection of prospective data, a different approach will be taken. The Consortium will draft a Research Protocol for the collection and processing of Prospective Data. The idea is to create a standardized informed consent form (GDPR and ethical consent), which will be handed out – and hopefully signed – by potential participants. The Consortium will comply with all legal and ethical obligations and recommendations.

5 ProCancer-I Open Data

5.1 Open Research Data Pilot

ProCancer-I project, is participating in the Open Research Data Pilot (ORDP) under Horizon 2020.

According to the EC[51]: *"the ORD pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects and takes into account the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservation questions. Participating in the ORDP does not necessarily mean opening up all research data. Rather, the ORDP follows the principle "as open as possible, as closed as necessary" and focuses on encouraging sound data management as an essential part of research best practice. The ORDP applies primarily to the data needed to validate the results presented in scientific publications. Other data can also be provided by the beneficiaries on a voluntary basis, as stated in their Data Management Plans."*

During the compilation of the second version of the current deliverable related to the Data Management Plan ProCancer-I consortium partners still examine which type of data collected and/or generated during the lifetime of the project will be public available after the end of the project.

More specifically, the following two issues will be explored in order the final decision to be made:

- **The data may be incompatible with the obligation to protect results** that can reasonably be expected to be commercially or industrially exploited.
- **The data may be incompatible with rules on protecting personal data** especially since in ProCancer-I raw and processed data from patients diagnosed with prostate cancer is foreseen.

5.2 Open access to scientific publications

Open Science means sharing knowledge and tools as early as possible, not only between researchers and between disciplines but also with society at large. Open access publishing (also called 'gold' open access), meaning that an article is immediately provided in open access mode by the scientific publisher, or self-archiving (also called 'green' open access) meaning that the published article or the final peer-reviewed manuscript is archived by the researcher- or a representative -in an online repository before, after or alongside its publication will be adopted. Authors copyrights agreements will determine whether scientific publications, resulted from the project, will adopt the gold or the green model. However, in the case copyright agreements are not violated (e.g. in the case of peer reviewed journals and international conference proceedings), the consortium will favor whichever model guarantees wider dissemination of the

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

project results. Publications will be uploaded on Zenodo[52] which helps researchers receive credit by making the research results citable and through OpenAIRE integrates them into existing reporting lines to funding agencies like the European Commission. Citation information is also passed to DataCite and onto the scholarly aggregators.

5.3 EU Recommendations on open data access

The following recommendations concerning open data access has been reported:

Commission's Recommendation on access to and preservation of scientific information of (C(2012) 4890 final), where EU Member State has nominated a National Point of Reference, with the task of reporting on the implementation of open access in the Member States. The 2012 Recommendation on access to and preservation of scientific information (2012/417/EU) was part of a package that outlined measures to improve access to scientific information produced in Europe and to bring them in line with the Commission's own policy for Horizon 2020[53].

Recommendation C(2018)2375, adopted on April 25th 2018[54], that explicitly reflects developments in areas such as research data management (including the concept of FAIR data i.e. data that is Findable, Accessible, Interoperable and Re-usable), Text and Data Mining (TDM) and technical standards that enable re-use incentive schemes. It reflects ongoing developments at the EU level of the European Open Science Cloud, and it more accurately takes into account the increased capacity of data analytics of today and its role in research. It also clearly identifies as two separate points the issue of reward systems for researchers to share data and commit to other open science practices on the one hand, and skills and competences of researchers and staff from research institutions on the other hand[55].

5.4 Link with European Research Infrastructures and the European Open Science Cloud

Special efforts will be devoted in making sure that seamless interoperability with other relevant European infrastructures will be achieved. In specific, interoperability with BBMRI-ERIC, that holds *omics* data which could in the future be fused with imaging data in developing advanced AI based decision support systems in the domain of PCa will be implemented. Several ProCAncer-I partners (CNR, FORTH) are involved in the technical Common Services IT (CSIT) working group of the BBMRI-ERIC and will guide the effort.

The European Open Science Cloud (EOSC) is Europe's virtual, federated, open environment in which the scientific community can access, store, manage, analyse and reuse digital research outputs (including publications, data, metadata and software) for research, innovation and educational purposes[56]. It provides cloud-based services for open sciences— integration and

[52] https://zenodo.org/
[53] http://ec.europa.eu/research/openscience/pdf/openaccess/background_note_open_access.pdf#view=fit&pagemode=none
[54] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0790
[55] http://ec.europa.eu/research/openscience/index.cfm
[56] https://www.eoscsecretariat.eu/sites/default/files/open_consultation_booklet_sria-eosc_20-july-2020.pdf

consolidation of e-infrastructure platforms, federation of existing European research infrastructures and scientific clouds.

The EOSC, funded through the Horizon 2020 initiative and officially launched in November 2018, has started providing access to services via their EOSC Portal[57]. ProCancer-I aims to provide resources through the EOSC portal. In order to assure compatibility in protocols and standards with the EOSC initiative we need to comply with the following requirements[58] :

- The resource is accessible by users outside its original community.
- The resource is described through a common template focused on value proposition and functional capabilities.
- At least one resource instance is running in a production environment available to the user community.
- Published Research data is Findable, Accessible, Interoperable and Reusable (FAIR).
- Release notes and sufficient documentation are available.
- Helpdesk channels are available for support, bug reporting and requirements gathering.

[57] Eudat, Liber, OpenAIRE, Egi, Geant, European Open Science Cloud for research [Internet], Position Paper, 2015 Oct. Available at http://libereurope.eu/wp-content/uploads/2015/11/OSC_Position_Paper-final-30.10.15.pdf
[58] https://providers.eosc-portal.eu/becomeAProvider

6 Risks

Any processing of personal data carried out by the controller or on the controller's behalf fall within the responsibility of the controller. In particular, the controller is obliged to implement appropriate and effective measures and be able to demonstrate the compliance of processing activities with the data protection regulations including the GDPR. The implemented measures should take into account the nature, scope, context and purposes of the processing and the risk to the rights and freedoms of natural persons (Recital 74 GDPR). These risks to the rights and freedoms of natural persons may result from personal data processing, which could lead to physical, material or non-material damage (Recital 75 GDPR).

D2.3 sets out the initial Data Protection Impact Assessment (DPIA). This deliverable identifies the potential risks to the rights and freedoms of the natural persons involved in processing of personal data in the ProCAncer-I project and the subsequent risk mitigation measures which should be observed by those partners processing personal data. The DPIA will be updated throughout the duration of the project when necessary.

7 Ethical aspects

Art. 19 of the Regulation establishing Horizon 2020[59] provides that any research and innovation activities carried out under Horizon 2020 shall not only comply with relevant national, Union and international legislation but also with the relevant ethical principles. Art. 34 of the ProCAncer-I Grant Agreement re-emphasizes the obligation that all beneficiaries need to comply with ethical and research integrity principles. Failing to incorporate these values would not only indicate irresponsible research that results in outputs of questionable value that may be seen as unreliable and high-risk, but would also constitute a breach of a beneficiary with the abovementioned obligations and may lead to significant adverse effects for the human subjects involved.

Therefore, the ProCAncer-I consortium considers ethics as an integral part of research from beginning to end, and ethical compliance is seen as pivotal to achieve real research excellence. Ethical compliance will furthermore facilitate public trust in the ProCAncer-I solution and increase credibility in the project's outputs.

While there are specific data protection and human rights laws that this project will fully adhere to ensuring the protection individuals, there are however no specific EU laws regarding ethics. In the words of the European Data Protection Supervisor: "Ethical thinking and deliberation come before, during, and after the law".[60] The fact that our research is legally permissible does not necessarily mean that it will be deemed ethical. Therefore, the responsibility lies within each participating partner to conduct their tasks in ways that include respect and protection of human values, which are intrinsic to the existing legislation and fully adhere to the highest ethical standards, as set out for example in The European Code of Conduct for Research Integrity.[61]

The primary ProCAncer-I objective is to shed light on the use of advanced AI techniques to exhaustively extract information from imaging regarding all the unmet clinical needs of prostate cancer patients management. As such, ProCAncer-I will involve research on data collected from patients with prostate cancer. All data collected, shared and analysed during the project will strictly follow the European legal and ethical regulations. ProCAncer-I focuses on creating a unique platform that will include the largest collection of PCa mpMRI images worldwide and will focus on novel AI clinical tools for advancing characterization of prostate lesions, assessing the metastatic potential, and early detection of disease recurrence. The exploitation of imaging and health data collected for therapeutic and diagnostic purposes within or without clinical trials on the one hand and health data collected for research purposes on the other raises ethical and

[59] Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020) and repealing Decision No 1982/2006/EC.
[60] https://edps.europa.eu/sites/edp/files/publication/19-03-25_reuters_interview_en.pdf
[61] ALLEA - All European Academies, The European Code of Research Integrity (2017) https://www.allea.org/wp-content/uploads/2017/05/ALLEA-European-Code-of-Conduct-for-Research-Integrity-2017.pdf

legal issues. For ethical reasons this will be addressed in the context of the patients' informed consent, particularly their right of withdrawal. From a legal point of view, this variation in the purpose of data use is generally prohibited. Therefore the ProCAncer-I project will adhere to the corresponding European legal and ethical regulations.

The consortium came to the conclusion that the informed consent of the subject with respect to data collected in the past ('retrospective data') will not be obtained subsequently, as it would significantly slow down the devolopement of ProstateNET and could cause emotional distress to former patients and their families. The legal and ethical reasoning for this approach can be found in the Retrospective Data Research Protocol[62]. This protocol will be amended by the local clinical partners to fit their national laws and guidelines and will be evaluated by local ethics committees. The retrospective datasets will be anonymized by the local clinical centres and as such from a data protection perspective, they will not be considered personal data anymore.

Prospective data will be collected to develop, test and validate the AI algorithms that the ProCAncer-I will deliver upon approval by local/national ethical review committees. For approval informed consent is mandatory. Collection of data will follow the rules of the countries in which it is collected. Informed consent will be obligatory in all prospective collection of data. All studies will be the subject of ethical reviews by the respective local ethical committees. The prospective data will also be fully anonymized, prior to transfer, and as such there will be no issue with data transfer across borders from a data protection perspective other than maintaining the anonymisation of that data. The patients will be informed about how there data is going to be used – even after the anonymisation. The information duties and the informed consent requirements for the prospective data collection will be determined in a Prospective Data research Protocol.

ProCAncer-I efforts to realize trustworthy AI, by adhering to EU guidelines on ethics in artificial intelligence,[63] will lead to the identification of new best practices / methods related to the design, development, and validation of AI solutions (e.g. privacy and data governance; transparency, diversity, non-discrimination and fairness; accountability). The ethical framework defined by IEEE[63] will serve as a reference for the ProCAncer-I consortium, aspiring to contribute to the realization of key requirements for achieving trustworthy AI: (i) the process models by which engineers and technologists can address ethical consideration throughout the various stages of system initiation, analysis and design, or (ii) specific methodologies addressing ethical values and principles (e.g. selection of data sets to eliminate bias, data privacy).[64]

The consortium will use state of the art anonoymization techniques in order to remove all identifiers, which could allow reidentification of a patient. However, to mitigate the remaining

[62] ProCancer-I retrospective protocol.
[63] https://standards.ieee.org/industry-connections/ec/autonomous-systems.html
[64] Geis JR et al. Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Radiology, 2019, 293.2: 436-440.

risk of unsufficient anonymisation, the access and tems of use of the data repository will be strictly regulated by the Terms & Conditions of the repository.

8 References

In the current deliverable the necessary references are introduced in the document as footnotes.

Appendix I – Dataset description template

The following table depicts the template format where the different datasets are described including valuable metadata for an efficient Data Management Monitoring.

Table: DS.#.# table template

Data identification: *DS#.#_* <dataset name>	
Generic description:	
<Provide a short description of the dataset>	
Origin of data:	
<Describe here the origination of the data that compile the specific dataset>	
Nature and scale of data:	
<file format of dataset>	
To whom the dataset could be useful:	
<Describe who could utilise/exploit the specific dataset>	
Related scientific publication(s)	
<Is the dataset related to a scientific publication? Is the latter Gold or Green Open Access?>	
Indicative existing similar data sets (including possibilities for integration and reuse):	
<Are there public available datasets similar to the specific one? If yes provide details.>	
Partners activities and responsibilities	
Partner owner of the data	-
Partner in charge of the data analysis	-
Partner in charge of the data storage	-
Related WP(s) and task(s)	-
Standards and metadata	
Standards, format, estimated volume of data	

	<Are there Standards that the dataset complies to? What is the dataset format? What is the estimated size of it?>
Data exploitation and sharing	
Data access policy/ Dissemination level: confidential (only for members of the Consortium and the Commission Services) or Public	<Classification of the Dissemination Level>
Data sharing, re-use, distribution, publication (How?)	-
Personal data protection: are they personal data? If so, have you gained (written) consent from data subjects to collect this information?	<Does the dataset include personal sensitive data?>
Access Procedures	-
Embargo periods (if any)	-
Archiving and preservation (including storage and backup)	
Data storage (including backup): where? For how long?	<Where are the data stored? What type of backup process is planned?
Indicative associated costs for data archiving and preservation	-
Indicative plan for covering the above costs	-

Appendix II – Tools and services for data FAIRification

Tool/Service	Reference	Description
Registry of Research Data Repositories	https://www.re3data.org/	re3data.org is a global registry of research data repositories for locating and accessing research datasets in all scientific disciplines. It provides an overview of existing research data repositories in order to help researchers to identify a suitable repository for their data and thus comply with requirements set out in data policies. re3data.org is now a regular service of DataCite.
Zenodo	https://zenodo.org/	Zenodo is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN. It allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artifacts. For each submission, a persistent digital object identifier (DOI) is minted and supports various data and license types. One supported source is GitHub repositories. Zenodo is compliant with the data management requirements of Horizon 2020 and Horizon Europe, the EU's research and innovation funding programmes.
FigShare	https://figshare.com/	Figshare is an online open access repository where researchers can share and preserve their research outputs, including figures, datasets, images, and videos. The files can be uploaded in any format and items are attributed a DOI. It is free to upload content and free to access, in adherence to the principle of open data. All files are released under a Creative Commons license. The main hosting mechanism for the platform is Amazon S3.which supports backup and preservation via a distributed cloud computing network.
Dryad	https://datadryad.org/stash	Dryad is an international open-access repository of research data, especially data underlying scientific and medical publications. Dryad is a curated general-purpose repository that makes data discoverable, freely reusable, and citable. Dryad serves as a repository for tables, spreadsheets, flat files, and all other kinds of published data for which

		specialized repositories do not already exist. All data files are are attributed a DOI and are made available for reuse under the terms of a Creative Commons Zero waiver. Dryad's metadata is supported by a Dublin Core metadata application profile.
FAIRfier (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/ https://github.com/FAIRDataTeam/Openrefine-metadata-extension	A general-purpose FAIRifier on the basis of the OpenRefine data cleaning and wrangling tool and the RDF plugin. This FAIRifier enables a post-hoc FAIRification workflow: load an existing dataset, perform data wrangling tasks, add FAIR attributes to the data, generate a linked data version of the data and, finally, push the result to an online FAIR data infrastructure to make it accessible and discoverable. Literal values in a dataset can be replaced by identifiers either manually or by embedded, customizable script expressions. The interoperability of the dataset can be improved by connecting these identifiers into a meaningful semantic graph-structure of ontological classes and properties using the integrated RDF model editor. A provenance trail automatically keeps track of each modification and additionally enables "undo" operations and repetition of operations on similar datasets. A FAIR data export function opens up a metadata editor to provide information about the dataset itself.
FAIR Data Point (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/	FAIR Data Point (FDP) is software that allows data owners to expose metadata and data in a FAIR manner. It offers a graphical user interface (GUI) for human clients and an application programming interface (API) for software clients. FDPs make datasets and their fine-grained metadata discoverable and accessible by machines. The datasets can be external or internal to the FAIR Data Point.
FAIR Search Engine (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/	The FAIR Data Search Engine harvests the metadata available on FAIR Data Points or compatible data repositories, indexes them, and provides a search interface

ORKA (part of Data FAIRport)	https://www.dtls.nl/fair-data/find-fair-data-tools/	The Open, Reusable Knowledge graph Annotator (ORKA) supports easy human curation of knowledge graphs by offering graph annotation as a service and capturing the provenance of the annotator and the original statement.
DataCite Metadata Creator	https://dhvlab.gwi.uni-muenchen.de/datacite-generator/ https://github.com/mpaluch/datacite-metadata-generator	A web based wizard for creating the XML metadata required by DataCite when registering a data DOI. Metadata is generated by populating text boxes and selecting values from drop-downs. The results can then be saved to a file.
FAIRsharing	https://fairsharing.org/	FAIRsharing is FAIRsharing is a web-based, searchable portal and FAIR-supporting resource that provides an informative and educational registry on data standards, databases, repositories and policy, alongside search and visualization tools and services that interoperate with other FAIR-enabling resources. FAIRsharing guides consumers to discover, select and use standards, databases, repositories and policy, and producers to make their resources more discoverable, more widely adopted and cited. Each record in FAIRsharing is curated in collaboration with the maintainers of the resource themselves. As of February 2019, FAIRsharing had over 2,620 records: 1,293 standards, 1,209 databases and 118 data policies. FAIRsharing collaborates with other infrastructure resources to cross-link each record to other registries, as well as within major FAIR-driven global initiatives, such as the FAIR Metrics working group (http://fairmetrics.org) The content within FAIRsharing is minted a DOI and licensed via the Creative Commons Attribution ShareAlike 4.0 license (CC BY-SA 4.0)
Center for Expanded	https://metadatacenter.org/	CEDAR is a platform for development, evaluation, use, and refinement of biomedical metadata aiming to

Data Annotation and Retrieval (CEDAR)		facilitate data discovery, data interpretation, and data reuse. The CEDAR Workbench provides a versatile, REST-based environment for authoring metadata that are enriched with terms from ontologies. The metadata are both stored in a CEDAR metadata repository and exported along with the primary data.
Castor EDC	https://www.castoredc.com/	Castor is a cloud-based clinical data management platform, enabling researchers to easily capture and integrate data from clinicians, patients, devices, wearables, and EHR systems.
Data Stewardship (DS) Wizard	https://ds-wizard.org/ https://github.com/ds-wizard	The DS Wizard is a tool for data management planning that is focused on getting the most value out of data management planning for the project itself rather than on fulfilling obligations. It is based on FAIR Data Stewardship, in which each data-related decision in a project acts to optimize the FAIRness of the data, explicitly guiding researchers in order to make their results FAIRer. The output of the DS Wizard is both a human-readable but also a machine-actionable, FAIR, Science Europe complaint, Data Stewardship Plan.