D3.7

Open Research Data Pilot

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	**v1**
Deliverable Type	Report
Distribution level	Public
Contractual Date of Delivery	30/09/2022
Actual Date of Delivery	3/10/2022

Authors	**Haridimos Kondylakis, Nikolaos Tachos, Manolis Tsiknakis**
Contributors	All WP Leaders
Reviewers	Nikos Papanikolaou, Daniele Regge

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
V0.1	Table of Contents & Structure	01.08.2022
V0.2	First version of the deliverable	01.09.2022
V0.3	Comments and reviews	15.09.2022
V0.4	Updates based on review comments	28.09.2022
V1.0	Final version (FORTH)	30.09.2022

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

This document contains material, which is the copyright of one or more ProCancer-I consortium parties, and may not be reproduced or copied without permission.

All ProCAncer-I consortium parties have agreed to full publication of this document.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the ProCAncer-I consortium as a whole, nor individual ProCancer-I consortium parties, warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, accepting no liability for loss or damage suffered by any person using this information.

Executive summary

D3.7 “Open Research Data Pilot” of the project ProCAncer-I is prepared under the activities of WP3 - Dissemination, Communication, Open data availability and Data management Plan. In line with the principles of Open Access to Research Data, ProCAncer-I will provide open access of outcomes and results of the project and enable reuse of knowledge. The Consortium is committed to participate in the Open research Data Pilot of the EU. Based on the initial version of the Data Management Plan will promote open access without violating the legal and ethical framework developed. Within this context, issues related to IPRs, personal data, sensitive information and compromise of project outcomes are continuously identified.

Table of Contents

List of Abbreviations

Abbreviation	Explanation
GA	Grant Agreement
ORDP	Open Research Data Pilot
DMP	Data Management Plan
GDPR	General Data Protection Regulation (EU) 2016/679
FAIR	Findability, Accessibility, Interoperability, Reusability

List of Tables

List of Figures

1. Introduction

ProCAncer-I's vision is to become a catalyst in this process by creating the first European, ethical- and GDPR compliant, quality-controlled, prostate cancer 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.

The Open Research Data Pilot in addition sets out detailed legal requirements on open access to scientific publications: under Horizon 2020, each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results. To meet this requirement the ProCAncer-I project will ensure that any scientific peer reviewed publications will be accessible online to be read, downloaded and printed.

Further the Open Research Data pilot establishes how the data will be stored openly and how shall be accessed. More specifically:

- Data collected through the lifetime of the project **will be available for free to the research community**, after submitting a justified application, which will be evaluated by a data access committee.
- **Selected data of the project will be published in AI challenges** (a first dataset has been released already to the PI-CAI (Prostate Imaging: Cancer AI) challenge[1] and two more datasets are planned to be released in the following hackathons that will organized by the project) **and also deposited in open research infrastructures** (e.g. Zenodo).

Further as data to be shared derive from real patients additional ethical and issues are raised which should be tackled whereas IPR rights on the collected data are of utmost importance.

The Consortium is committed to participate in the Open research Data Pilot of the EU. Based on the initial version of the Data Management Plan will promote open access without violating the legal and ethical framework developed. In essence the D3.7 "Open Research Data Pilot" of the project ProCAncer-I provides a first plan on how outcomes and results of the project will be openly made available and enable reuse of knowledge, and is a sequel of the first version of the data management plant (D3.5). As the project evolve the presented plan will evolve and gain more precision and substance and will be reported in the subsequent versions of the data management plan and will be concluded with the final data management plan (D3.6).

1.1 Deliverable structure

The current deliverable is structured in the following sections. In Section 2 we report on the open access plan established for publications. Then in Section 3 we elaborate on the types of data that

[1] https://pi-cai.grand-challenge.org/

will be collected and generated through the lifetime of the project and we elaborate on the plan for making them public available. Finally, Section 4 elaborates on the risks and IPR rights.

2. Open access to publications

Open access (OA) refers to the practice of providing online access to scientific information that is free of charge to the end-user and reusable Open access to scientific publications means free online access for any user. Although there are no legally binding definitions of 'access' or 'open access' in this context, authoritative definitions of open access appear in key political declarations including:

- the 2002 Budapest Declaration
- the 2003 Berlin Declaration

Under these definitions, 'access' includes not only basic elements - the right to read, download and print – but also the right to copy, distribute, search, link, crawl and mine. The 2 main routes to open access are:

- **Self-archiving / 'green' open access** – the author, or a representative, archives (deposits) the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication. Some publishers request that open access be granted only after an embargo period has elapsed.
- **Open access publishing / 'gold' open access** - an article is immediately published in open access mode. In this model, the payment of publication costs is shifted away from subscribing readers. The most common business model is based on one-off payments by authors. These costs, often referred to as Article Processing Charges (APCs) are usually borne by the researcher's university or research institute or the agency funding the research. In other cases, the costs of open access publishing are covered by subsidies or other funding models.

In the context of research funding, open access requirements do not imply an obligation to publish results. The decision to publish is entirely up to the grant beneficiaries. Open access becomes an issue only if publication is chosen as a means of dissemination.

To meet this requirement, beneficiaries must, at the very least, ensure that any scientific peer reviewed publications can be read online, downloaded and printed.

Since any further rights -such as the right to copy, distribute, search, link, crawl and mine - make publications more useful, beneficiaries should make every effort to provide as many of these options as possible.

The open access mandate comprises 2 steps:

a) depositing publications in repositories

b) providing open access to them

2.1 Repositories

All ProCAncer-I partners will deposit a machine-readable electronic copy of the published version or final peer-reviewed manuscript accepted for publication in a repository for scientific publications. This must be done as soon as possible and at the latest upon publication (The latest acceptable time to deposit a publication is the date of publication). Repository' for scientific publications is an online archive. Institutional, subject-based and centralized repositories are all acceptable choices. Repositories that claim rights over deposited publications and preclude access are not.

The Open Access Infrastructure for Research in Europe (OpenAIRE) is the recommended entry point for researchers to determine what repository to choose.

2.2 Providing open access

After depositing publications beneficiaries will ensure open access to those publications via the chosen repository. Beneficiaries can choose one of two main ways to meet this requirement:

- **Self-archiving / 'green' OA:** beneficiaries can deposit the final peer-reviewed manuscript in a repository of their choice. They must ensure open access to the publication within at most 6 months (12 months for publications in the social sciences and humanities). To provide support concerning compliance with Horizon 2020 embargo periods the Commission offers a model amendment to publishing agreement, which are often signed between authors and publishers. This model is not mandatory but reflects the obligations for the beneficiary under the H2020 grant agreements. It can be supplemented by further provisions agreed between the parties, provided they are compatible with the Grant Agreement. The Commission/Agency takes no responsibility for the use of this model.
- **Open access publishing / 'gold' OA**: researchers can also publish in open access journals, or in hybrid journals that both sell subscriptions and offer the option of making individual articles openly accessible. Monographs can also be published either on a purely open access basis or using a hybrid business model. 'Article processing charges' are eligible for reimbursement during the duration of the project (as other costs defined in article 6.2.D.3 of the Model Grant Agreement). As stated, the article must also be made accessible through a repository upon publication.

The costs of 'gold' open access publications incurred once a project is completed cannot be refunded from that project's budget. Beneficiaries must also provide open access, through the repository, to the bibliographic metadata that identify the deposited publication. These must be in a standard format and must include the following:

- terms ["European Union (EU)" & "Horizon 2020"]["Euratom" & Euratom research & training programme 2014-2018"]
- name of the action, acronym & grant number
- publication date, the length of the embargo period (if applicable) and a persistent identifier.

3. Open Access to Data & Data Altruism

The Commission has enabled access to and reuse of research data generated by Horizon 2020 projects through the Open Research Data Pilot (ORD Pilot).

Types of data covered by the Open Research Data Pilot

- 'Underlying data' (the data needed to validate the results presented in scientific publications), including the associated metadata (i.e. metadata describing the research data deposited), as soon as possible
- Any other data (for instance curated data not directly attributable to a publication, or raw data), including the associated metadata, as specified and within the deadlines laid down in the DMP – that is, according to the individual judgement by each project/grantee.

Requirements of the Open Research Data Pilot

Step 1: The project must **deposit the research data** described above, preferably in a research data repository. These are online research data archives, which may be subject-based/thematic, institutional or centralized. Useful listings of repositories include the Registry of Research Data Repositories and Databib.

ProCAncer-I will store all datasets collected in ProstateNET, a European research data repository formulated and sustained by the project. Further **selected datasets will be published in OpenAIRE's repository, Zenodo[2],** an OpenAIRE and CERN collaboration, which allows researchers to deposit both publications and data, while providing tools to link them. Zenodo and some other repositories as well as many academic publishers also facilitate linking publications and underlying data through persistent identifiers and data citations. Datasets released include the ones released for AI challenges in cancer imaging (e.g. PIC-AI) or hackathons organized by the project.

Step 2: As far as possible, projects must then take measures to **enable third parties to access, mine, exploit, reproduce and disseminate** (free of charge for any user) this research data. One straightforward and effective way of doing this is to attach Creative Commons Licences (CC BY or CC0) to the data deposited. The EUDAT B2SHARE tool includes a built-in license wizard that facilitates the selection of adequate license for research data.

ProCAncer-I will enable third parties to access and exploit the available data through a regulated data governance model that will be established in the next phases of the project. More specifically the datasets will be accessible upon written request by research organizations whereas they will also be accessible from companies subject to a small fee that will enable the sustainability of the platform in the future. The specific fee structure and if any will be decided in the following phases of the project. Nevertheless, the project coordinator, FORTH, has opened an account in ZENODO and offers to all project beneficiaries the possibility to publish selected research data openly with the coordinator's account or to create their own accounts.

[2] https://zenodo.org/

3.1 Data Summary

As per D3.5, for self-completeness, we present in the sequel the dataset that will be collected the lifetime of the project per WP

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)

For further details on the aforementioned datasets and the methodology for data collection the interested reader is referred to D3.5.

4. Ethical aspects and intellectual property rights

4.1 Ethical Issues

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

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

The ethical aspects of the Project are continuously assessed under WP9, and more specifically D9.1 and D9.2 were already successfully submitted, setting out the 'ethics requirements' that the project must comply with.

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

4.2 Confidentiality & Safeguarding

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

The partners have considered the following additional safeguards in accordance with Article 89 GDPR:

- **Data minimization** will be respected in scientific research conducted within ProCAncer-I. This means that only necessary data will be processed. The respect of the data minimisation principle is ensured by the methodology used to define the necessary data needed for the project. The data collection protocols and templates were drafted based on input provided by all partners, taking into account ProCAncer-I's AI developers and use cases. The partners have extensive experience on which they can build, including in prior projects, as well as documented good practices to use as reference. Ethical Committees have approved the relevant protocols, further substantiating compliance with the proportionality requirement.
- **Data de-identification**. The responsible partners developed data anonymization protocol and guidelines. Moreover, data import and anonymization tools were made available to the Data Providers, who are also responsible for submission of the Data.

- **DPIA**: When there is a large-scale processing of sensitive personal data or when that processing is likely to result in a high risk for the data subject, a data protection impact assessment is needed. The Article 29 Working Party has recommended that a DPIA be performed as well when the processing entails storage for archiving purpose of pseudonymized personal data concerning vulnerable data subjects of research projects or clinical trials. Therefore, a DPIA was already submitted at M6 (D2.3).
- **Purposes of scientific processing:** The consortium partners are developing a Joint Controller Agreement committing to specified and defined purposes of use of the data within the ProCAncer-I project, and after its completion. The JCA is will be signed prior to the completion of the project and before public access is to be given to the data assets collected.
- **National law.** Additional safeguards may be imposed through national law as well. Member States can, for instance, define the modalities of anonymization, or the procedures for obtaining ethics clearance. Detailed conditions imposed by the national laws were examined and described in DPIA for Retrospective data.
- **Ethical approvals.** Ethical approval actions for data collection were completed.
- **Data access tracking**. Due to security requirements, the access is regulated by several measures. The data are securely stored centrally in the ProstateNet.
- **Assignment of processing roles.** The Parteners have signed data sharing agreement and data processing agreement in which the roles and responsibilities of have been assigned.
- **Controlled access to data:** The consortium will develop a data governance model on the basis of controlled access to data in the Hybrid repository.
- **Ensuring IT security through**:
    - Storage of data in the Central node
    - Secure infrastructure / space
    - Encryption / Virtual Private Network (VPN)
    - Logging access to data / Transaction Tracker
    - Security against attacks within the operating system / Hardware Security
    - Identity Access Management
    - Data access control mechanism
    - Protection of AI models
    - Data cataloguing, indexing and retrieval mechanism

4.3 IPR

Issues regarding the protection of intellectual property rights (IPRs) and confidential information were addressed in detail within the Consortium Agreement. In particular, the Consortium Agreement regulates the IP-Ownership, Access Rights to Background and Foreground IP (Articles 8, 9 and 10). Moreover, in accordance with Article 24 of the Grand Agreement, Background was identified for all Partners, if applicable. The details will be described in the Joint/Composite IPR agreement for the exploitation of ProCancer-I AI models.

5. Conclusions

The document presented the initial ProCAncer-I Open Research Pilot plan following the initial Data Management Plan already submitted and available. This deliverable established the plan for openly disseminating publications and data and also elaborated on ethical concerns and IPR issues. The ProCAncer-I DMP will be revised and updated during the entire duration of the project and finalized in D3.6.