D4.4

Initial version of the ProCAncer-I platform

Related Work Package	**WP4 – Imaging repositories, sharing mechanisms, annotation, curation, and standardization methodologies**
Related Task	Task 4.3 - Design and implementation of imaging repositories Task 4.4 - Image and data annotation tool Task 4.5 - Sharing and curation tools Task 4.6 - Implement monitoring, logging, and retraining of AI Models Task 4.7 - High performance computing & services for AI pipelines
Lead Beneficiary	ADVANTIS
Contributing Beneficiaries	B3D, FORTH, QUIBIM, CNR
Document version	**v1.0**
Deliverable Type	Demonstrator
Distribution level	Public
Contractual Date of Delivery	30/11/2021
Actual Date of Delivery	24/12/2021

Authors	**Christos Pollalis (Advantis)**
Contributors	Stelios Sfakianakis, Valia Kalokyri, Eugenia Mylona (FORTH) Joao Correia, Walter Hernandez (B3D) Ana Jimenez Pastor (QUIBIM)
Reviewers	Haridimos Kondylakis [FORTH], Nikolaos Tachos [FORTH]

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	ToC	01/11/2021
v0.2	Contributions by FORTH	02/11/2021
v0.3	Contributions by QUIBIM, B3D	12/11/2021
v0.4	Contributions by QUIBIM, B3D, FORTH, ADVANTIS	18/11/2021
v0.5	1st revision of the deliverable	29/11/2021
v0.6	Annexes updates and revisions [FORTH]	14/12/2021
v0.7	Updates in the section 5.3 [CFO]	16/12/2021
v0.8	Updates in the Section 5 of the deliverable [FORTH]	17/12/2021
v0.9	Revisions to all deliverable sections [FORTH]	20/12/2021
V1.0	Final Version – Upload to EC	24/12/2021

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This deliverable contains original unpublished work except where clearly indicated otherwise.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

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

ProCAncer-I aims to create a scalable, sustainable, quality-controlled, prostate-related, medical imaging platform, where large-scale data and Artificial Intelligence (AI) algorithms will co-exist. The platform will promote multi-center interoperability via a large multiparametric (mp) Magnetic Resonance Imaging (MRI) data repository alongside tools for data analysis and sharing. Advanced processing steps will be utilized aiming at a deeper understanding of imaging findings.

This document outlines the initial version ProCAncer-I Platform's architecture and describes its various storage implementations, services and tools, AI models' framework, and monitoring and logging systems. At the current point of project's development, a lot of emphasis is given to the upload of the retrospective data and its constituents processes: the data preparation in accordance with the ethical and security related requirements (anonymization), the annotation and curation using domain specific tools and metadata standards, and their efficient storage management to support scalable retrieval of structured information. In more detail, this document focuses on the following aspects of the architecture:

- The data upload process from the end user's point of view and the eCRF application. We briefly describe the selected "use cases", i.e., the patient case classification to answer the clinical questions that are the primary focus of the project;
- The data annotation and curation tools that are used for imaging related tasks like segmentation, motion correction, and co-registration;
- The core data management components, i.e., the DICOM Image Repository and the Metadata Repository, and their crucial role in the data upload process;
- The security related services to support users and services authentication and authorization and their integration with the ELIXIR federated identity infrastructure for cross-organization single sign-on.

Furthermore, this document provides the initial design for the ProCAncer-I Machine Learning/AI framework that will allow the deployment of state-of-the-art model development approaches ("MLops") supporting the whole model lifecycle: from model building and tuning with "experiment tracking", to model deployment and serving, and to monitoring of the "real world" performance and the alerting mechanisms to address challenges such as data and concept drifts.

The whole infrastructure is currently deployed on a commercial cloud but in a "cloud agnostic" way to avoid vendor lock-in problems and facilitate porting to a "cloud native" but provider independent solution in the future version

Table of Contents

List of Abbreviations

Abbreviation	Explanation
AAI	Authentication & Authorization Infrastructure
ABAC	Attributed Based Access Control
AI	Artificial Intelligence
AMQP	Advanced Message Queueing Protocol
API	Application Programming Interface
AuthN	Authentication
AuthZ	Authorization
AWS	Amazon Web Services
CDM	Common Data Model
CNCF	Cloud Native Computing Foundation
DICOM	Digital Imaging and Communications in Medicine
eCRF	Electronic Clinical Report Form
ETL	Extract, Transform, and Load
FAIR	Findability, Accessibility, Interoperability, Reusability
GCP	Google Cloud Platform
IdP	Identity Provider
JSON	JavaScript Object Notation
JWT	JSON Web Token
ML	Machine Learning
mpMRI	Multi-Parameter Magnetic Resonance Imaging
OMOP	Observational Medical Outcomes Partnership
OPA	Open Policy Agent
PACS	Picture Archiving and Communications System
PCa	Prostate Cancer
PDP	Policy Decision Point
PEM	Privacy Enhanced Mail
PEP	Policy Enforcement Point
PHI	Protected Health Information
PI-RADS	Prostate Imaging Reporting & Data System
QIDO-RS	Query based on ID for DICOM Objects by RESTful Services
RBAC	Role Based Access Control

Abbreviation	Explanation
RDBMS	Relational Database Management System
STOW-RS	Store Over the Web by RESTful Services
UI	User Interface
WADO-RS	Web Access to DICOM Objects by RESTful Services

List of Tables

List of Figures

1. Introduction

This document is the fourth Deliverable of WP4: Imaging Repositories, Sharing Mechanisms, Annotation, Curation, and Standardization Methodologies, comprises.

This document outlines the ProCAncer-I Platform's architecture and describes its various storage implementations, services and tools, AI models' framework, and monitoring and logging systems. Its content is based on T4.1: Definition of Platform Architecture - AI Pipelines, whose outcome is documented in D4.1: Report on requirements elicitation and initial design of the platform.

In order to provide the clinical context of the work in ProCAncer-I to the readers of the present document and because we have not documented the clinical Uses Cases (UCs) that drive the ProCAncer-I vision in any of the Deliverables of the project that have already been delivered, we have decided to include a short Chapter (Chapter 2) in which to describe the main clinical questions we seek to answer by exploiting the data and remaining services that comprise the ProCAncer-I platform.

Subsequently, Chapter 3 gives an overview of the ProCAncer-I Platform's architecture and the technologies of the underlying infrastructure.

Chapter 4 describes the various storage systems: the long-term DICOM store, the temporary/staging area of uploaded studies, the accompanying meta-data catalogue, also including curation information, and the clinical data repository.

Chapter 5 describes all (internal and user-facing) platform services: authentication and authorization backend and middleware, and user tools for data upload, sharing, curation, and annotation, as well as implementation details.

Chapter 6 describes the ecosystem of tools implemented/integrated to develop, (re-)train, and manage the AI models and monitor their performance.

Chapter 7 describes continuous, platform-wide monitoring and logging solutions.

This document details the alpha version of the ProCAncer-I platform. It verifies the availability of the first versions of the high-performance computing infrastructure, imaging repository, meta-data catalogue, as well as sharing, curation, annotation, and visualization tools. Proper authentication, authorization, and anonymization mechanisms are ready in order to securely enable upload of the retrospective data.

2. ProCAncer-I clinical use-cases

Nowadays, prostate cancer is highly curable if detected in its early stages. Indeed, when organ-confined at diagnosis, the five-year survival rate of prostate cancer is close to 100%. Unfortunately, current prostate cancer diagnostic pipeline has several limitations. Prostate specific antigen (PSA) has a low sensitivity and specificity. Indeed, approximately 15% of men with normal PSA values (below 4 ng/ml) will have prostate cancer, while in 75% of subject with PSA above 4 ng/ml the only findings will be prostate hypertrophy or inflammation. Moreover, when high values of PSA are detected in the blood of patients, biopsy is necessary to confirm diagnosis. Unfortunately, up to 30% of clinically significant tumours will be missed if a systematic approach is adopted for biopsy, i.e. sampling is performed randomly under ultrasound guidance.

When these tests fail to identify prostate cancer, patient's long-term health and well-being might suffer an irreversible impact, because the later the diagnosis, the more likely that cancer will have grown and spread to remote parts of the body becoming a deadly disease. Hence, the main clinical challenges for the future will be to identify prostate cancer with high accuracy, as early as possible, to stratify patients according to disease aggressiveness and to tailor therapy (or non-therapy) based on the risk of progression, comorbidities and life expectancy.

In this context, the Use Cases within the ProCAncer-I project represent the drivers to answer PCa relevant clinical questions, ranging from PCa diagnosis and characterization to prediction of treatment response and occurrence of side effects after treatment. The Use Cases were defined to create a unique dataset in terms of data quantity, quality and diversity, and to facilitate validation of the platform with its added value and usability for various users among the clinical and research community.

The DoA of the project describes the development of AI validated models for a range of pressing clinical scenarios and questions that include: 1) Accurate detection of prostate cancer both in the peripheral as well as the transitional zone; 2) Characterization of cancer according to its biological aggressiveness into clinically significant and non-significant disease; 3) Identification of patients with metastatic prostate cancer as early as possible; 4) Radiologic – Histopathologic correlation to provide biology-based validation of AI models; 5) Prediction of the risk of disease recurrence; 6) Prediction of treatment response in case of Radiation therapy; 7) Prediction of post radical prostatectomy and/or Radiation-induced urinary toxicity; 8) AI-powered patient stratification for enrolment in active surveillance programs. These are the 8 Use Cases (UCs) in which the project is focusing that cover the whole disease continuum. During the initial phases of project implementation an additional Use Case was discussed and it was agreed to be included into the work programme of the project.

Therefore, the 9 Use Cases that constitute the focus of the work in ProCAncer-I are:

- **UC1:** Detection of prostate cancer with high accuracy both in peripheral and transitional zone to identify which men have cancer and those with no cancer. From a clinical point of view UC1 will help stratifying men with prostate hypertrophy or inflammation despite the high PSA values (>4 ng/ml) and those who should undergo additional diagnostic tests (e.g. biopsy) to identify if suspicious lesions identified on MRI are clinically significant or if there is an indolent disease with no harm for the patient;

- **UC2:** Characterization of cancer according to its biological aggressiveness into clinically significant and non-significant disease. UC2 aims to stratify men with suspicious findings on MRI into high-risk cases, which need radical treatments to ensure that cancer will not grow and spread to remote parts of the body becoming a deadly disease, from low-risk cases which could be safely follow-up with active surveillance, avoiding comorbidities of treatment and ensuring the highest possible quality of life for patients;
- **UC3:** Identification of patients with metastatic prostate cancer as early as possible among cases with high-risk PCa. Clinically, UC3 will help to adjust treatment strategies and mitigate metastatic spread that will finally kill the patient, as well as adjusting follow up frequency to patients with high metastatic risk. AI models will provide early indications whether a patient belongs to the metastatic subtype that needs a different therapeutic approach, mining tumor characteristics that probably are related to its biological differences;
- **UC4:** Radiologic – Histopathologic correlation to provide biology-based validation of AI models to compare side by side pathologic data with AI results to improve understanding of the features that AI models are making use to reach specific decisions. Moreover, UC4 will help correlating imaging phenotype derived from MRI to microscopic findings from pathology and predicting cancer presence and/or its biology characteristics from radiologic imaging;
- **UC5:** Prediction of the risk of disease recurrence after radical prostatectomy, based on imaging data and AI techniques. In UC5 post-surgery findings, as positive surgical margins and extracapsular extension, will be considered in a nomogram comprising also radiomics and clinical variables to predict disease recurrence. UC5 will help clinicians to choose between different treatment techniques (conventional, nerve sparing, laparoscopic, robot-assisted radical prostatectomy), tailoring treatment to the predicted risk of disease recurrence;
- **UC6:** Prediction of treatment response in case of radiation therapy, assessing the risk of disease recurrence to promptly adjust therapeutic strategy at an early stage and avoid patient discomfort and non-optimal distribution of medical resources. UC6 is similar to UC5, but refers radiotherapy recurrence and it will help radiation oncologist to tailor treatments;
- **UC7:** Prediction of post radical prostatectomy and/or radiation-induced urinary toxicity, in order to consider additional or alternative measures to alleviate therapy-induced undesired effects. Using UC7 results, patients with high risk for toxicity can be thoroughly informed on the side effects and alternative possibilities for therapy. This could help balancing the risk-benefit ratio related to whole gland treatments, in particular in patients with no life-threatening PCa. UC7 will take into consideration urinary incontinence, irritative/obstructive bowel, sexual/erectile dysfunction, and hormonal domains;

- **UC8:** AI-powered patient stratification for enrolment in Active Surveillance programs, to develop a more efficient patient stratification program based on AI decision-making from MRI lesion phenotype. The risk of disease progression in patients who are undergoing active surveillance will be assess with longitudinal MRI data (combined with biopsy) to reach specific clinical indications (either repeat PSA test, MRI, biopsy or a combination of them). UC8 aims also to predict the time-to-progress to provide a follow-up strategy and stratify patients in those who could safely remain in the active surveillance group and those who will ultimately need treatment.
- **UC9:** Prediction of the best option for patients needing treatment ensuring the lowest possible side effects/toxicity. Results from all previous Use Cases are expected to merge into a holistic model suggesting presence/non-presence of PCa, stratification into clinically significant/insignificant cases and a decision support system suggesting the best treatment option (radical prostatectomy, radiation therapy, or active surveillance), considering also the lowest toxicity/side effects to ensure the best possible quality of life.

In such a clinical context the data to selected prepared and uploaded into the ProCAncer-I technical platform for analysis and model development, per Use Case, is shown in the subsequent figure.

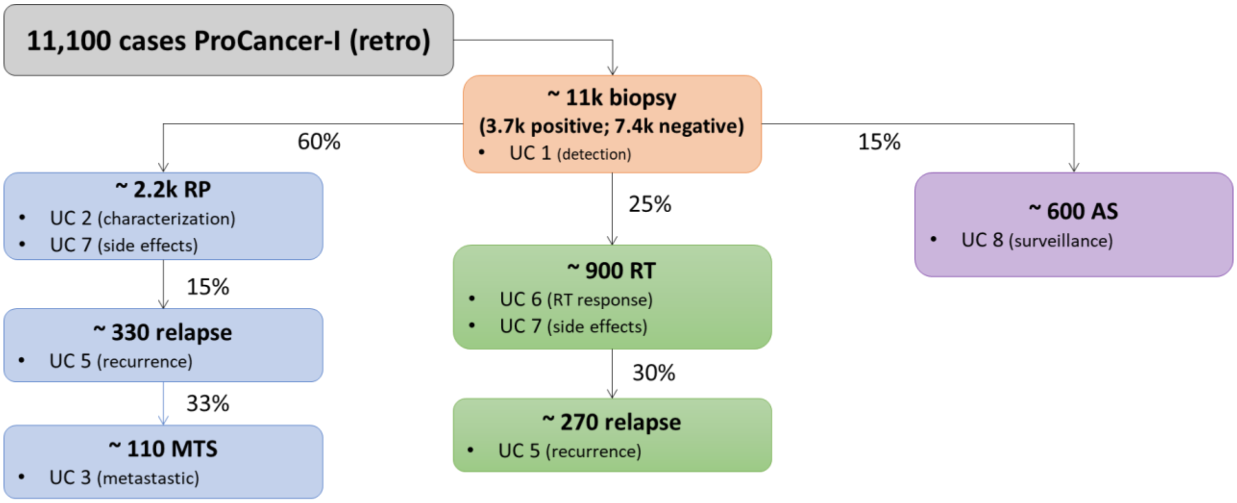


Figure 1. Summary of the estimated number of subjects included in each single use case[1].

[1] UC = use case; RP = radical prostatectomy; RT = radiation therapy; AS = active surveillance; MTS = metastatic disease

3. ProCAncer-I Platform Architecture

This section provides an overview of the ProCAncer-I Platform's architecture and underlying infrastructure. It summarizes the contents of D4.4: initial design of the platform and outlines the design decisions behind Task 4.7: High performance computing & services for AI pipelines, leading to an elastic, scalable, and fault-tolerant architecture.

As presented in Figure 2 the pipeline includes the steps of Data Collection, Data Preparation, the AI model Training & Validation and the Model Deployment. All these are describing the basic ML operations (MLOps) needed which are supported by the ProCAncer-I cloud platform. In the initial, alpha version of the platform described in the present deliverable (D4.4) the needed tools and services required to support data collection, and data preparation, including anonymization, upload, and curation, are designed, and developed along with the basic services and tools for the monitoring of the AI model training and validation.

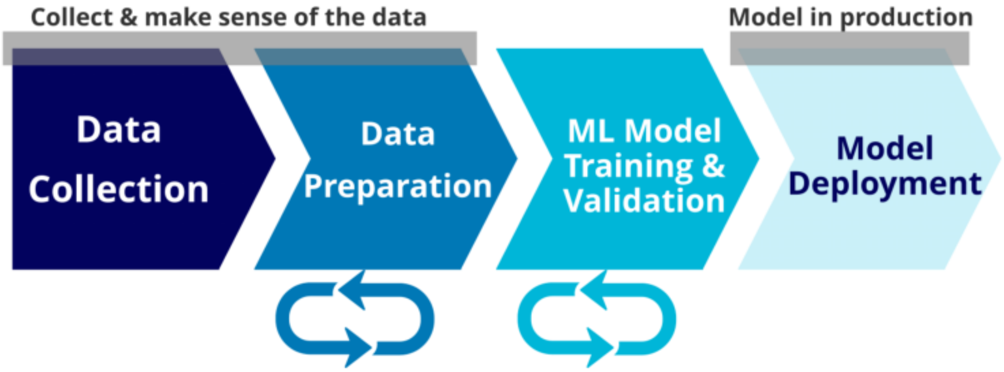


Figure 2. Conceptual pipeline of the ProCAncer-I development stages from the data collection to the AI models in production.

3.1 Overview

ProCAncer-I aims to deliver an infrastructure that follows the principles of open source, FAIR data access, common look-n-feel, common authentication and authorization, layered developing of modelling service, modelling service certification and cloud infrastructure independence, as stated in *D4.1 – Report on requirements elicitation and initial design of the platform*.

The logical view of the ProCAncer-I platform with the main domain specific areas of functionality of the system is shown in Figure 3 below.

The following subsystems are identified:

- *Data ingestion and upload.* This includes all the infrastructure (tools, services, cloud resources) that allows a data provider to upload their data sets according to the project's guidelines and best practices (e.g. anonymization) so that they become integrated to the curated cancer-related data managed by the system.

- *Data Management*, which supports the "data at rest" scenarios, is the core of the platform supporting all the other subsystems for the storage, efficient indexing, curation, and retrieval of the data.
- *Domain specific tools,* for example image and data annotation and data *tools*, which support domain experts to annotate and curate the imaging data.
- *Model management*. This is the part of the platform supporting the management of computational and AI tools and models. It allows searching for available models, the development of new ones, model execution and monitoring, etc.
- *Data and Service "Peering"* tools, that support the exchange of data and interoperability of services with other research infrastructures using well defined FAIR-enabled APIs and applications like the "Honest Broker".

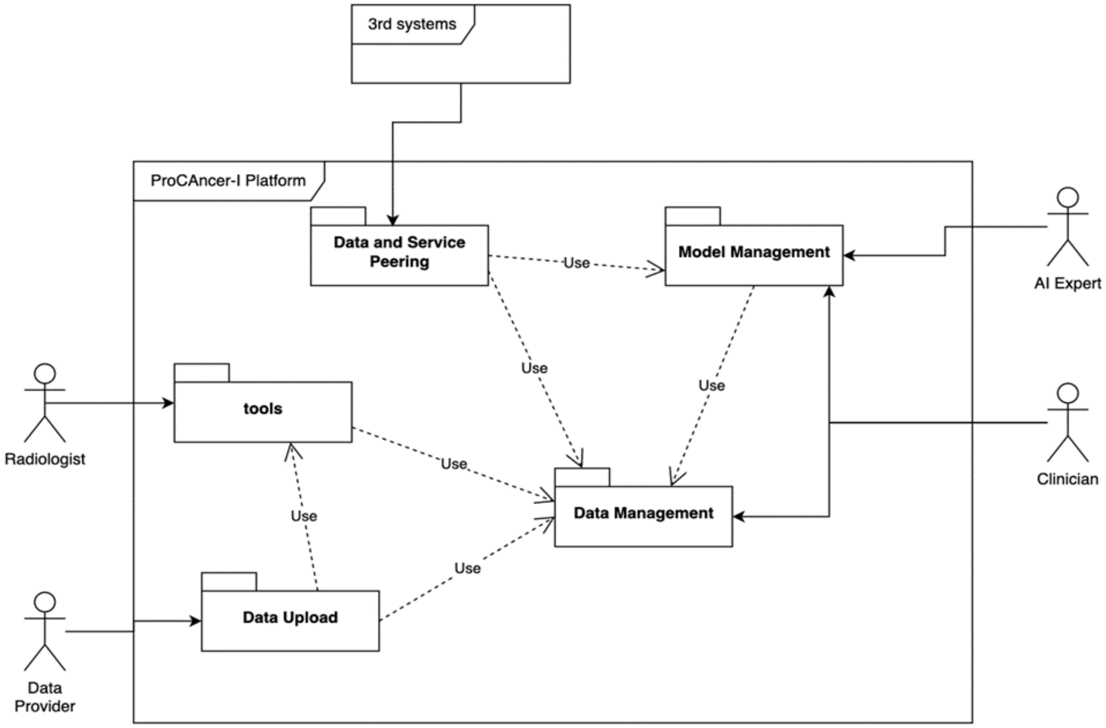


Figure 3. The main subsystems of the ProCAncer-I platform

The implementation of the platform is aligned with the Gantt Chart of the project. Hence, during the initial implementation period the main focus has been on the design, development and delivery of the infrastructure and tools to enable data collection and its preparation, including de-identification, for upload into the platform. Hence, data ingestion and upload, data management and domain specific tools have been developed and their integration has been concluded to allow data providers to make their data sets available to the ProCAncer-I community securely and fully annotated. These systems and tools are presented in this document. The remaining subsystems of the integrated platform, regarding model training, model management and data and service peering are in the development phase. These will be reported in the next |deliverable D4.5 - |final version of the ProCAncer-I platform, interoperability and AI services due on Month 24 of the project life-cycle.

3.2 Cloud Infrastructure

The cloud infrastructure deployed so far supports the required imaging and clinical repositories, and curation, visualization and segmentation services to start the process of data acquisition. The current version of the cloud infrastructure is based on Microsoft Azure cloud services. The ProCAncer-I platform at the current stage is based on "cloud-enabled" applications, that is applications that are built and tested "on premises" but deployed as virtual machines in the Azure cloud. The roadmap for the technical deployment of the platform concludes with the provision of "cloud native" functionality, that is specifically designed for the cloud, deployed and fixed faster, with a fluid architecture. In the near future, the replacement of the virtual machines with containers, which embody loosely-coupled microservices managed by a container orchestrator such as Kubernetes[2], is an important step to this direction and will improve the horizontal scaling of the platform and the "elasticity" to varying work load.

At any step in this evolutionary process, the platform will remain "cloud agnostic" i.e., portable and not rely on any specific cloud's APIs. Such a development approach requires additional effort by the project's technical partners but there is an evolving and fast-growing ecosystem of open source, cloud-native tools and services under the auspices of Cloud Native Computing Foundation (CNCF) to support this process[3].

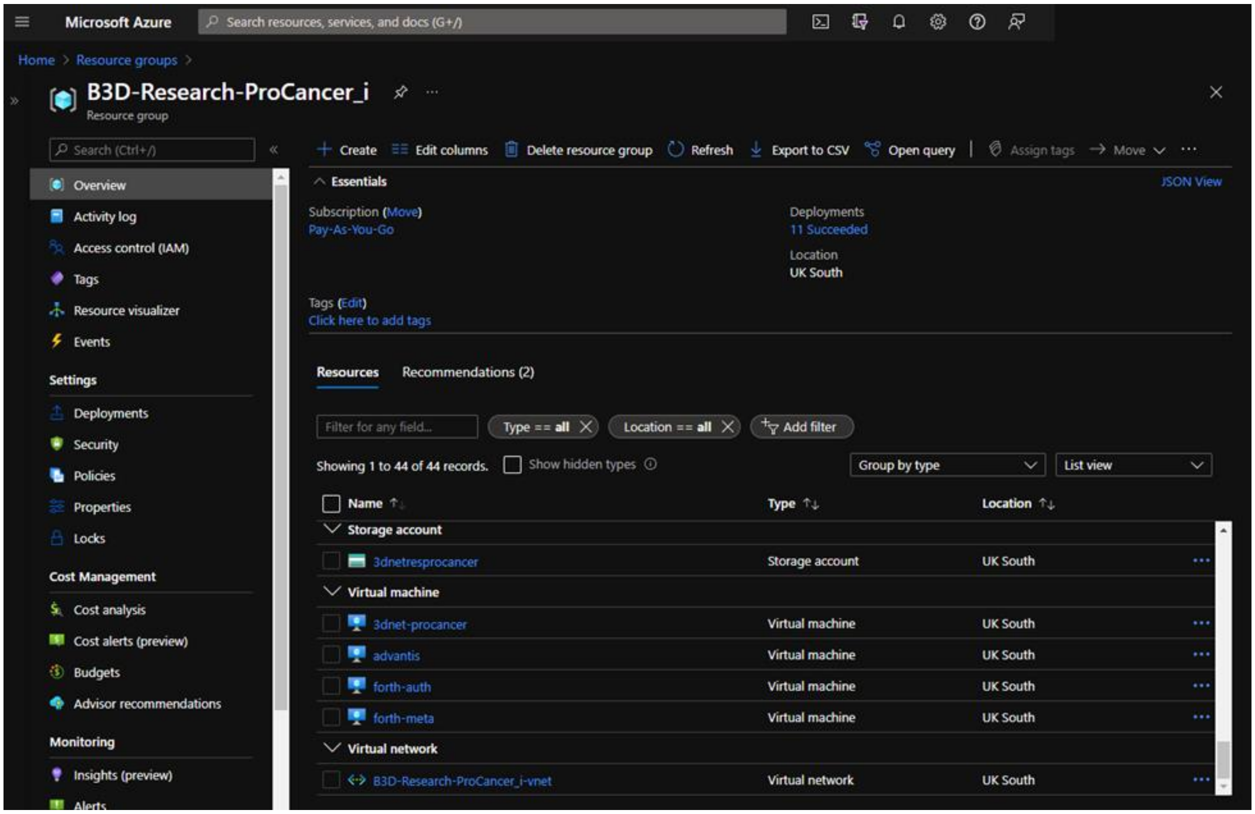
Taking into account the needs of the different services, in the current version of the platform an appropriate number of virtual machines have been deployed by Biotronics3D – the cloud infrastructure provider that are managed by the relevant project partner delivering the services:

- 3Dnet image repository managed by B3D (MS Windows server)
- Authentication server managed by FORTH (Ubuntu Linux)
- Clinical Data repository and Meta-data Catalogue server managed by FORTH (Ubuntu Linux)
- Image quality assurance server managed by ADVANTIS (Debian Linux)
- Image segmentation server managed by QUIBIM (Ubuntu Linux)

Through the Azure web-view which is a high-level management view, shown in Figure 4, Biotronics3D monitors and manages the ProCAncer-I cloud infrastructure.

[2] "Principles of Container-based Application Design" https://kubernetes.io/blog/2018/03/principles-of-container-app-design/

[3] https://www.cncf.io/ , https://landscape.cncf.io/



B3D-Research-ProCancer_i
Resource group

Essentials

Subscription (Move)
Pay-As-You-Go

Deployments
11 Succeeded

Location
UK South

Tags (Edit)
Click here to add tags

Showing 1 to 44 of 44 records.

Name	Type	Location
Storage account		
3dnetresprocancer	Storage account	UK South
Virtual machine		
3dnet-procancer	Virtual machine	UK South
advantis	Virtual machine	UK South
forth-auth	Virtual machine	UK South
forth-meta	Virtual machine	UK South
Virtual network		
B3D-Research-ProCancer_i-vnet	Virtual network	UK South

Figure 4. High-level management view in Azure web view

The partners involved in WP4 have been developing and integrating the cloud services and the local data upload tools to enable the upload and annotation of retrospective exams.

The ProCAncer-I cloud solution is supported on cloud infrastructure services, able to be deployed in different cloud service providers without substantial modifications, consistent with the requirement for being a “cloud-agnostic” platform. The following version of the cloud infrastructure will support container orchestration based on open frameworks. Kubernetes is an open-source container orchestration engine for automating deployment, scaling, and management of containerized applications. The ProCAncer-I aims to use Kubernetes considering the major cloud infrastructure providers, Amazon AWS, Microsoft Azure and Google Cloud Platform (GCP), offer Kubernetes Services.

4. ProCAncer-I Platform Storage

The ProCAncer-I platform will collect and manage large amounts of multimodal data (mpMRI imaging data and related clinical data) and metadata to be used for the training of advanced AI models in an efficient and clinically oriented fashion for prostate cancer management. The ProCAncer-I cloud platform storage, ProstateNet, is comprised of 3 components: the DICOM Object Store which stores medical imaging data; the clinical data document store which stores the clinical data; and the meta-data catalogue which stores metadata and semantic annotations to enable rich search and discovery of data and its exploitation. The flow of data is illustrated in Figure 5. The clinical partners will use a local, integrated eCRF and data upload tool to organise the DICOM studies and complete the clinical information, validate the use case, anonymise data and upload data to the cloud staging area. Each Clinical Partner will be able to run the data curation tools (if needed), verify the anonymisation and completeness of data, and submit validated cases to the ProstateNet repository (so called "staging area").

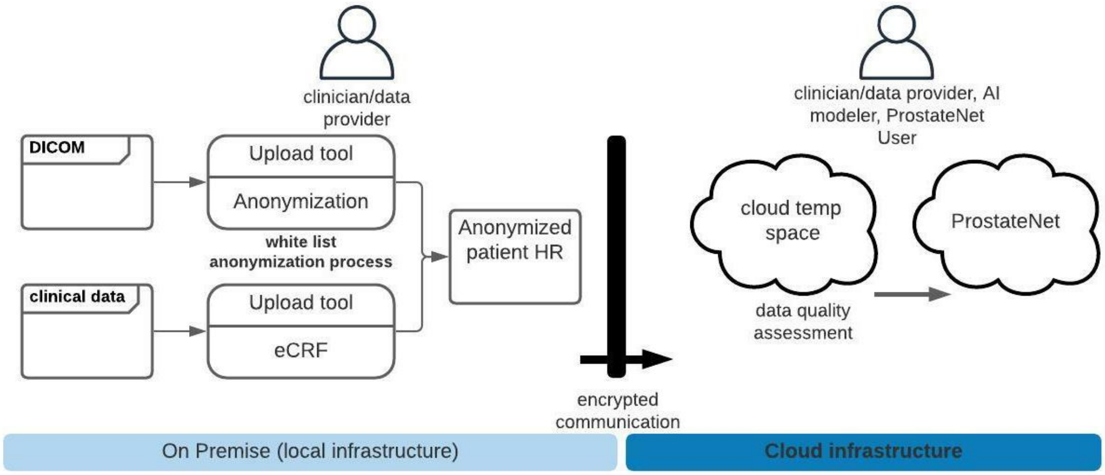
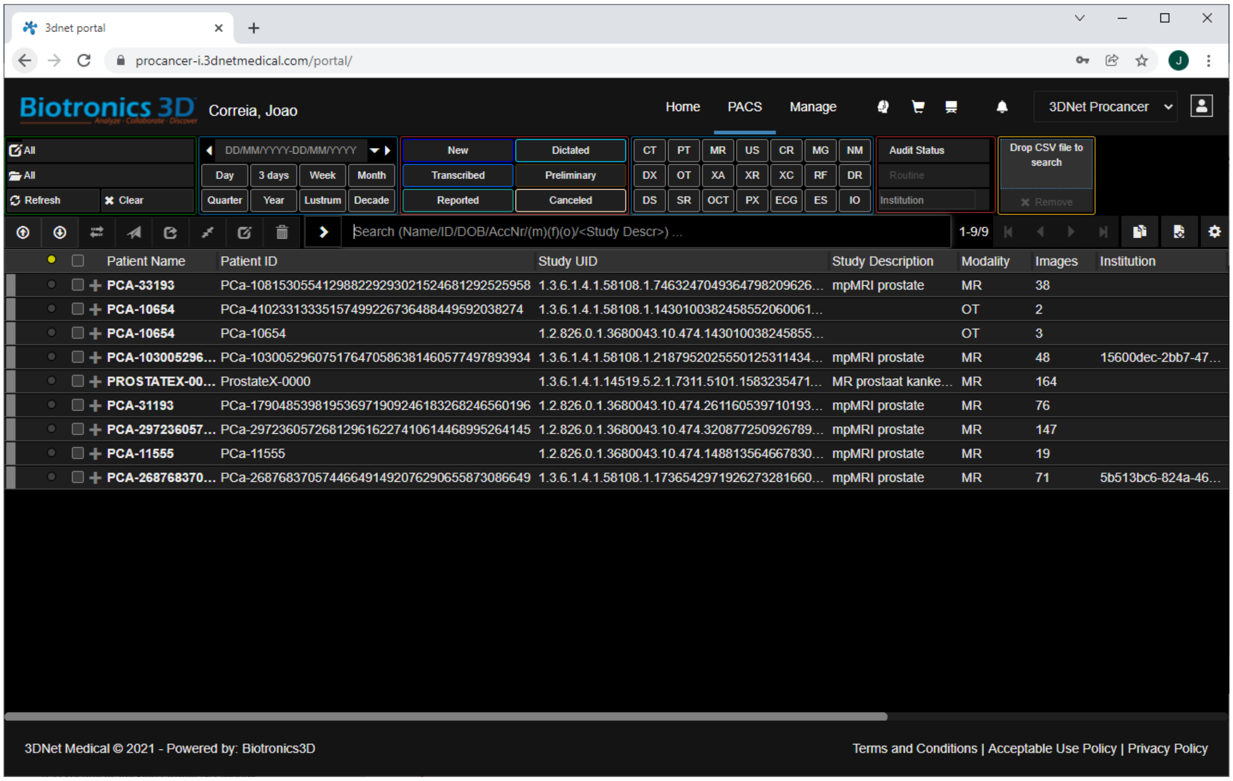


Figure 5. Data acquisition pipeline

4.1 DICOM Store

The ProCAncer-I project DICOM Object Store is the central DICOM object repository of the platform, which provides the necessary services for saving, updating, and retrieving DICOM studies. The implementation of the repository is based on the 3Dnet cloud medical imaging solution, from Biotronics3D, which is compliant with both the DICOM and HL7 standards, thus allowing seamless interoperability with existing PACS systems and scanners. Besides robust security, multi-organisation and roles-based user management, data will be anonymised before upload to prevent any disclosure of personal data. The repository is deployed at the Biotronics3D data centre, hosted at Microsoft Azure ISO 27001 accredited data centres – providing security, redundancy, reliability and scalability. To support the several steps of data curation, annotation and AI research and development, the repository provides mechanisms for querying and retrieving data through the API gateway. The user interface provides study browsing (Figure 6) and study and image visualization (Figure 7 and Figure 8) modules that are being integrated with the curation and annotation tools.

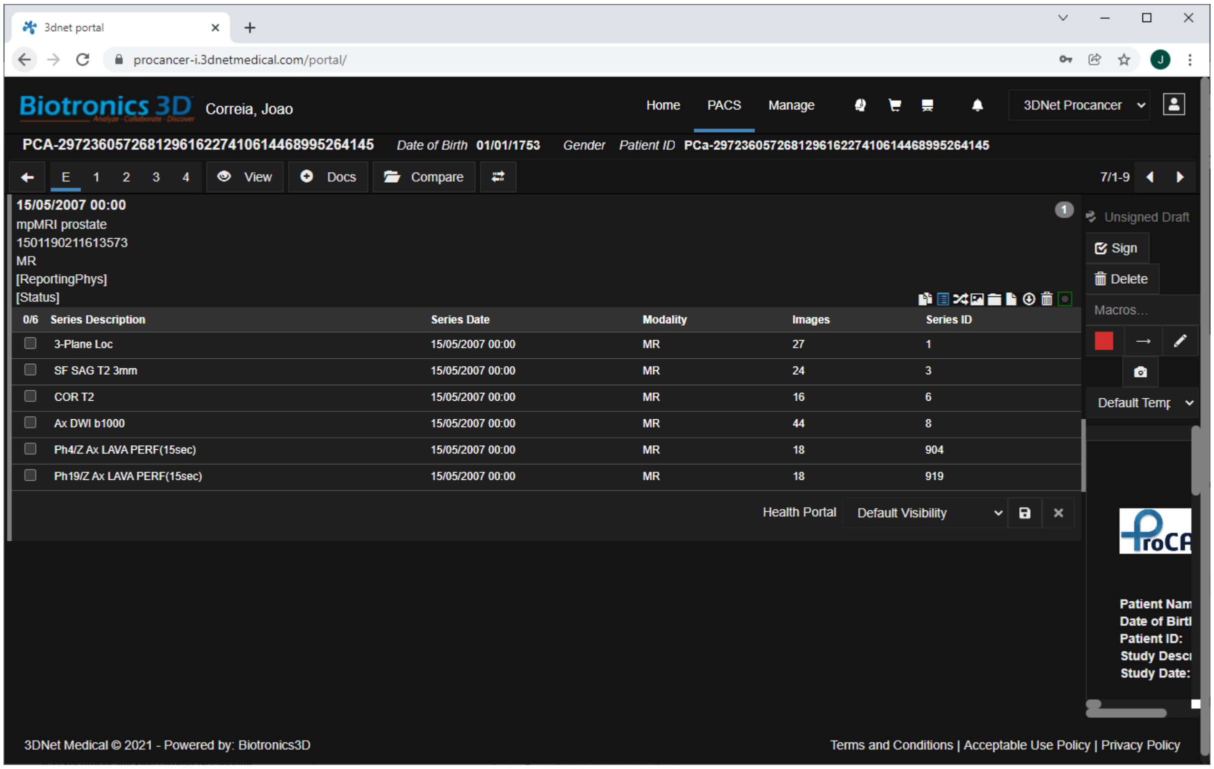


Patient Name	Patient ID	Study UID	Study Description	Modality	Images	Institution
PCA-33193	PCa-108153055412988229293021524681292525958	1.3.6.1.4.1.58108.1.7463247049364798209626...	mpMRI prostate	MR	38	
PCA-10654	PCa-41023313335157499226736488449592038274	1.3.6.1.4.1.58108.1.1430100382458552060061...		OT	2	
PCA-10654	PCa-10654	1.2.826.0.1.3680043.10.474.143010038245855...		OT	3	
PCA-103005296...	PCa-103005296075176470586381460577497893934	1.3.6.1.4.1.58108.1.2187952025550125311434...	mpMRI prostate	MR	48	15600dec-2bb7-47...
PROSTATEX-00...	ProstateX-0000	1.3.6.1.4.1.14519.5.2.1.7311.5101.1583235471...	MR prostaat kanke...	MR	164	
PCA-31193	PCa-179048539819536971909246183268246560196	1.2.826.0.1.3680043.10.474.261160539710193...	mpMRI prostate	MR	76	
PCA-297236057...	PCa-297236057268129616227410614468995264145	1.2.826.0.1.3680043.10.474.320877250926789...	mpMRI prostate	MR	147	
PCA-11555	PCa-11555	1.2.826.0.1.3680043.10.474.148813564667830...	mpMRI prostate	MR	19	
PCA-268768370...	PCa-268768370574466491492076290655873086649	1.3.6.1.4.1.58108.1.1736542971926273281660...	mpMRI prostate	MR	71	5b513bc6-824a-46...

3DNet Medical © 2021 - Powered by: Biotronics3D

Terms and Conditions | Acceptable Use Policy | Privacy Policy

Figure 6. Study browser of the DICOM object store showing test data



PCA-297236057268129616227410614468995264145 *Date of Birth* 01/01/1753 *Gender* *Patient ID* PCa-297236057268129616227410614468995264145

15/05/2007 00:00
mpMRI prostate
1501190211613573
MR
[ReportingPhys]
[Status]

0/6	Series Description	Series Date	Modality	Images	Series ID
□	3-Plane Loc	15/05/2007 00:00	MR	27	1
□	SF SAG T2 3mm	15/05/2007 00:00	MR	24	3
□	COR T2	15/05/2007 00:00	MR	16	6
□	Ax DWI b1000	15/05/2007 00:00	MR	44	8
□	Ph4/Z Ax LAVA PERF(15sec)	15/05/2007 00:00	MR	18	904
□	Ph19/Z Ax LAVA PERF(15sec)	15/05/2007 00:00	MR	18	919

3DNet Medical © 2021 - Powered by: Biotronics3D

Terms and Conditions | Acceptable Use Policy | Privacy Policy

Figure 7: Study Viewer of the DICOM object store showing test data

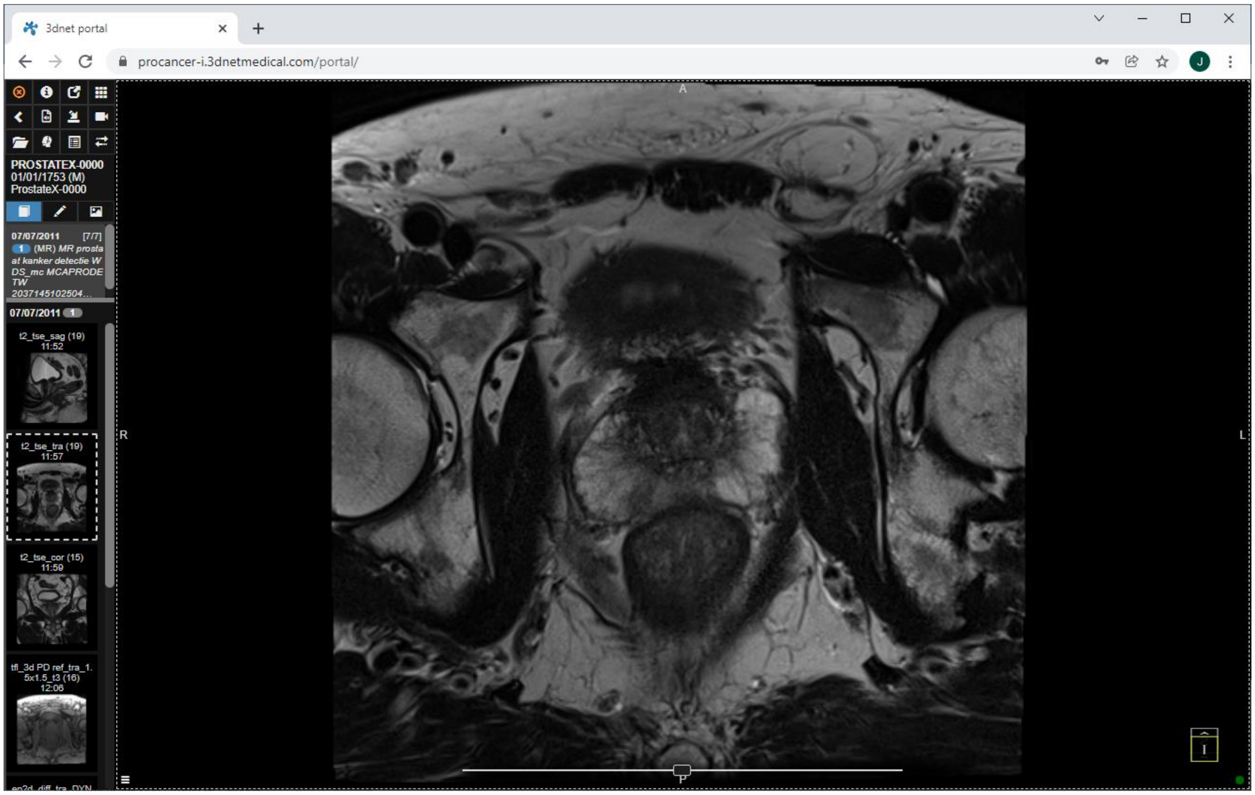
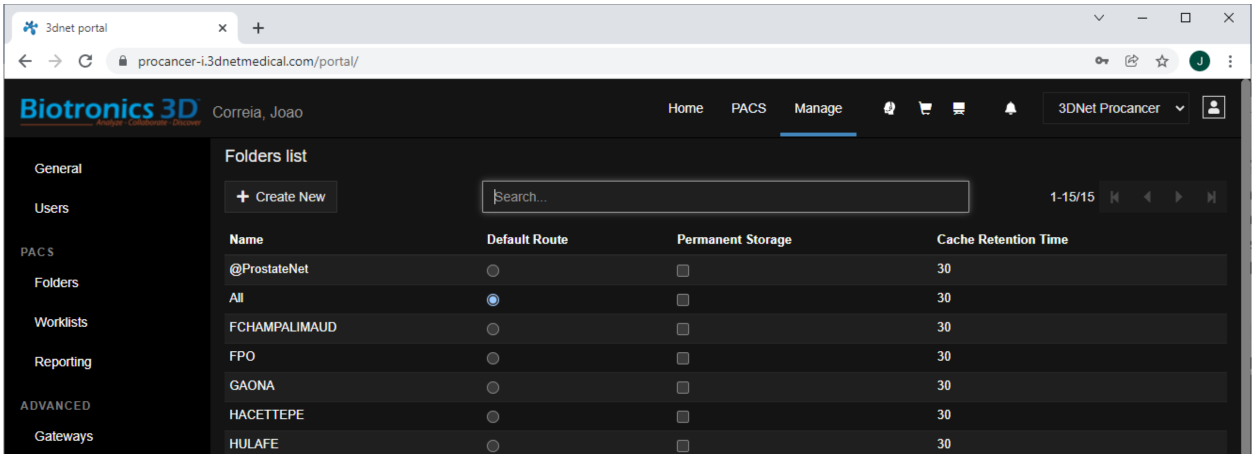


Figure 8. Image Viewer of the DICOM object store showing test data

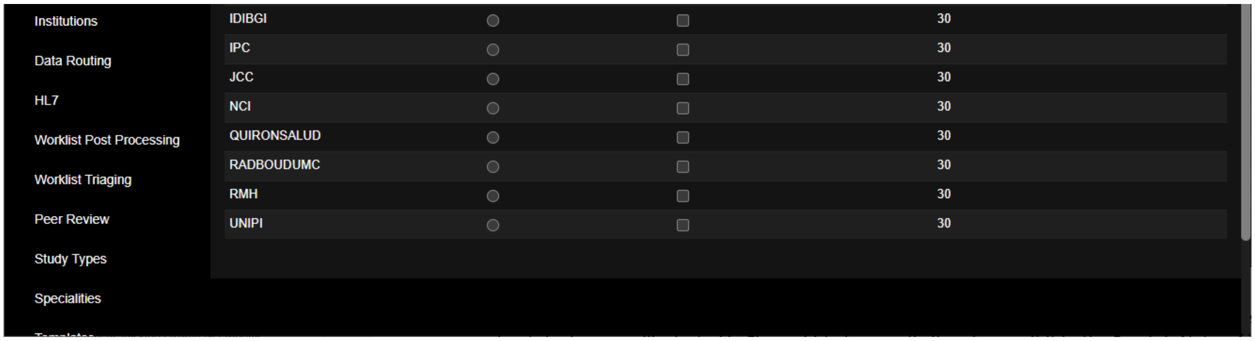
4.1.1. Staging Area

Each ProCAncer-I Clinical Partner has a special DICOM Object Store area where in which they can upload their new data. This area enables users to run data curation tools (if necessary) and verify data quality and anonymisation requirements. After the verification assessment, the user can transfer the DICOM studies to the main ProCAncer-I DICOM repository - ProstateNet.



Folders list

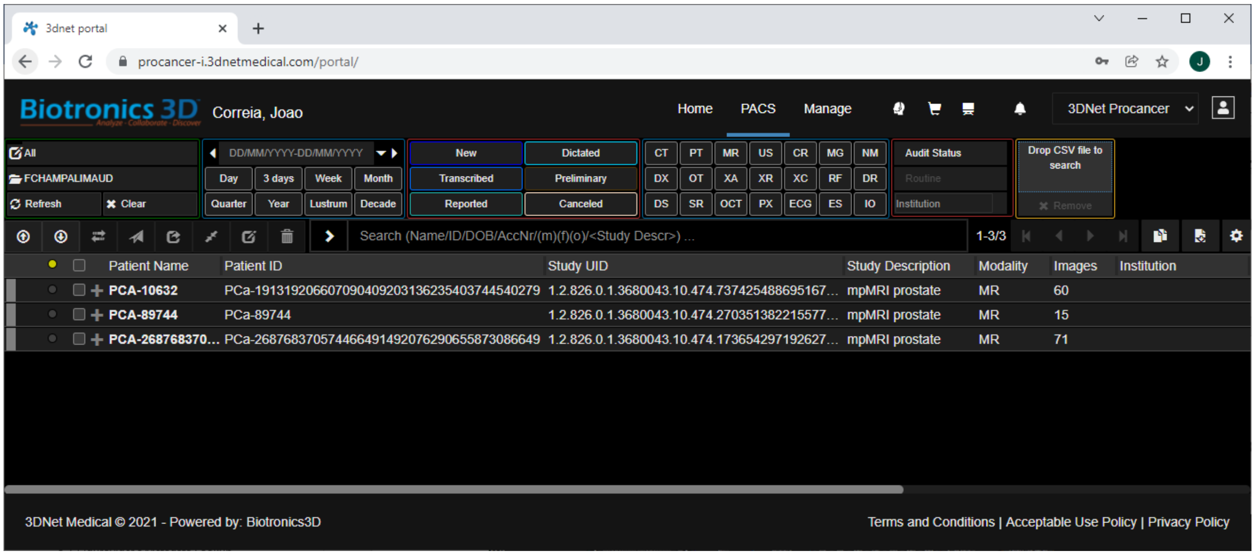
Name	Default Route	Permanent Storage	Cache Retention Time
@ProstateNet	○	□	30
All	●	□	30
FCHAMPALIMAUD	○	□	30
FPO	○	□	30
GAONA	○	□	30
HACETTEPE	○	□	30
HULAFE	○	□	30



IDIBGI	○	□	30
IPC	○	□	30
JCC	○	□	30
NCI	○	□	30
QUIRONSALUD	○	□	30
RADBOUDUMC	○	□	30
RMH	○	□	30
UNIPI	○	□	30

Figure 9. Staging areas in ProCAncer-I platform

After uploading a new case, the users of the Clinical Partner will be able to access the case on the cloud staging area, illustrated in Figure 10 below, verify it and commit it to the ProstateNet main repository.



Patient Name	Patient ID	Study UID	Study Description	Modality	Images	Institution
PCA-10632	PCa-191319206607090409203136235403744540279	1.2.826.0.1.3680043.10.474.737425488695167...	mpMRI prostate	MR	60	
PCA-89744	PCa-89744	1.2.826.0.1.3680043.10.474.270351382215577...	mpMRI prostate	MR	15	
PCA-268768370...	PCa-268768370574466491492076290655873086649	1.2.826.0.1.3680043.10.474.173654297192627...	mpMRI prostate	MR	71	

Figure 10. Staging area of a Clinical Partner showing 3 test studies

4.2 Clinical Data Repository

The Clinical Data Repository is the central data warehouse of the ProCAncer-I platform. In the data upload process this repository stores the submitted clinical and image related data before the metadata extraction and their persistence in the Metadata Catalogue. In this way, the data ingestion follows an Extract-Load-Transform (ELT) design where this repository is responsible for maintaining the data in their original submitted format. In more detail it offers the following features:

- It is the main repository of the clinical data uploaded to the platform. The metadata repository offers a similar role but it is oriented to a more organized and standards compliant (using OMOP) view of the information. Instead, the Clinical Repository is similar to a "data lake" or in a more accurate terminology to a "lake house"[4] that contains all of the uploaded information (except the imaging data) in the format that was uploaded. Being similar to a

[4] "What Is a Lakehouse?" https://databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

traditional data warehouse it offers transactional interactions and both structured and unstructured data storage.

- It contains imaging related metadata, for example selected DICOM tags extracted from the uploaded data, in a quasi-relational schema so that complex queries are possible and efficient. In contrast to the Imaging Repository that offers a mostly key-based "blob" storage (e.g. retrieving a DICOM series by its series UID), the Clinical Repository is able to cope with lots of different search criteria and access patterns.
- It stores information about the results of the image processing tools (e.g. segmentations) and their parameters allowing the linking between imaging data and provenance related metadata extraction.
- It maintains an upload log so to enable traceability, by gathering of statistics and monitoring of the use of the data platform are possible.

The current implementation of the repository is based on the PostgreSQL open source relational data management system. In this setup, PostgreSQL is used both as a classical RDBMS and as a document store: the initial data upload uses the JSON format and these data are stored as "JSONB" (binary JSON) in the database. PostgreSQL allows the creation of indexes over the JSON data for speeding up queries, but we also have defined "materialized" views where the same data are normalized to support easier querying and retrieval.

For the rest of the platform, the Repository is accessible through a web-based ("RESTful") API. In the initial version of the platform, this API is quite minimal offering a feed of the most recent uploads using the "JSON feed"[5] syndication format, accessing each patient data upload through its patient or upload id, and the retrieval of the patient data according to the "use case" that have been classified to. In the future, the API will be enriched with more functionality based on the requirements of the models and the other components of the platform.

4.2.1 Meta-Data Catalogue

In ProCancer-I, the MOLGENIS[6] platform has been adopted to serve as the main metadata catalogue of the project, whereas OMOP-CDM[7] and its extensions are used as the common data model to store the available metadata.

MOLGENIS is a modular web application for scientific data. It has been used for biobanking, rare disease research, patient registries and energy research. It enables capturing, exchanging and exploiting. It has a completely customized data system allowing modeling of the data using external data models. This creates flexibility that other, more static, database applications often lack. In addition, it is modular, having several modules to store and interact with the stored data, and provides interfaces to create R and Python scripts that interact with the data. MOLGENIS takes away the hassle of storing data, and makes it highly accessible with filters and fast search capabilities.

[5] "JSON Feed" https://www.jsonfeed.org

[6] https://www.molgenis.org/

[7] https://www.ohdsi.org/data-standardization/the-common-data-model/

Data Models. ProCAncer-I adopts the OMOP-CDM, which is one of the most widely used common data models for supporting analysis of observational health data, to support the generation of reliable scientific evidence about disease history, effects of medical interventions and health care interventions and outcomes. Besides the standard CDM, OMOP-CDM extensions are used, such as the Oncology CDM extension for representing cancer data at the levels of granularity and abstraction required to support cancer research. For radiology exams, although those can be currently registered using the OMOP-CDM, the model does not enable the storage of the subsequent curation process. As such, the ProCancer-I aspires to introduce a radiology extension and is currently working on it in collaboration with the OHDSI Medical Imaging Working Group[8], focusing on including annotation, segmentation and curation data as radiomics features that need to be stored as well.

OMOP-CDM ETL. To get from the native/raw data provided by the clinical sites (ProCAncer-I clinical partners) to the OMOP Common Data Model (CDM) an extract, transform, and load (ETL) process was defined and implemented. This process transforms the data from its initial raw format to the CDM, and adds mappings to a set of Standardized Vocabularies. The full ETL document can be found in the Appendix A. Terms found in the source data are mapped to concepts in the OMOP standard vocabularies to achieve semantic interoperability. In most cases a mapping to a standard concept with the same meaning as the source term can be made. If this is not possible, the source term is mapped to a non-standard concept. If a non-standard concept matching the source term does not exist either, then we create a custom 'ProCAncerI' concept. The ProCAncer-I custom vocabulary can be found in the Appendix A - Table 2.

ProCAncer-I aspires to exploit state of the art AI techniques (Radiomics and Deep Learning) to develop AI validated models for a range of pressing clinical scenarios and questions that include, as presented in Chapter 2:

1) Accurate detection of prostate cancer both in the peripheral as well as the transitional zone;
2) Characterization of cancer according to its biological aggressiveness into clinically significant and non-significant disease;
3) Identification of patients with metastatic prostate cancer as early as possible;
4) Radiologic – Histopathologic correlation to provide biology-based validation of AI models;
5) Prediction of the risk of disease recurrence;
6) Prediction of treatment response in case of Radiation therapy;
7) Prediction of post radical prostatectomy and/or Radiation-induced urinary toxicity;
8) AI-powered patient stratification for enrolment in active surveillance programs;
9) Prediction of the best treatment with lowest side effects/toxicity.

[8] https://forums.ohdsi.org/t/new-working-group-medical-image-working-group/15098

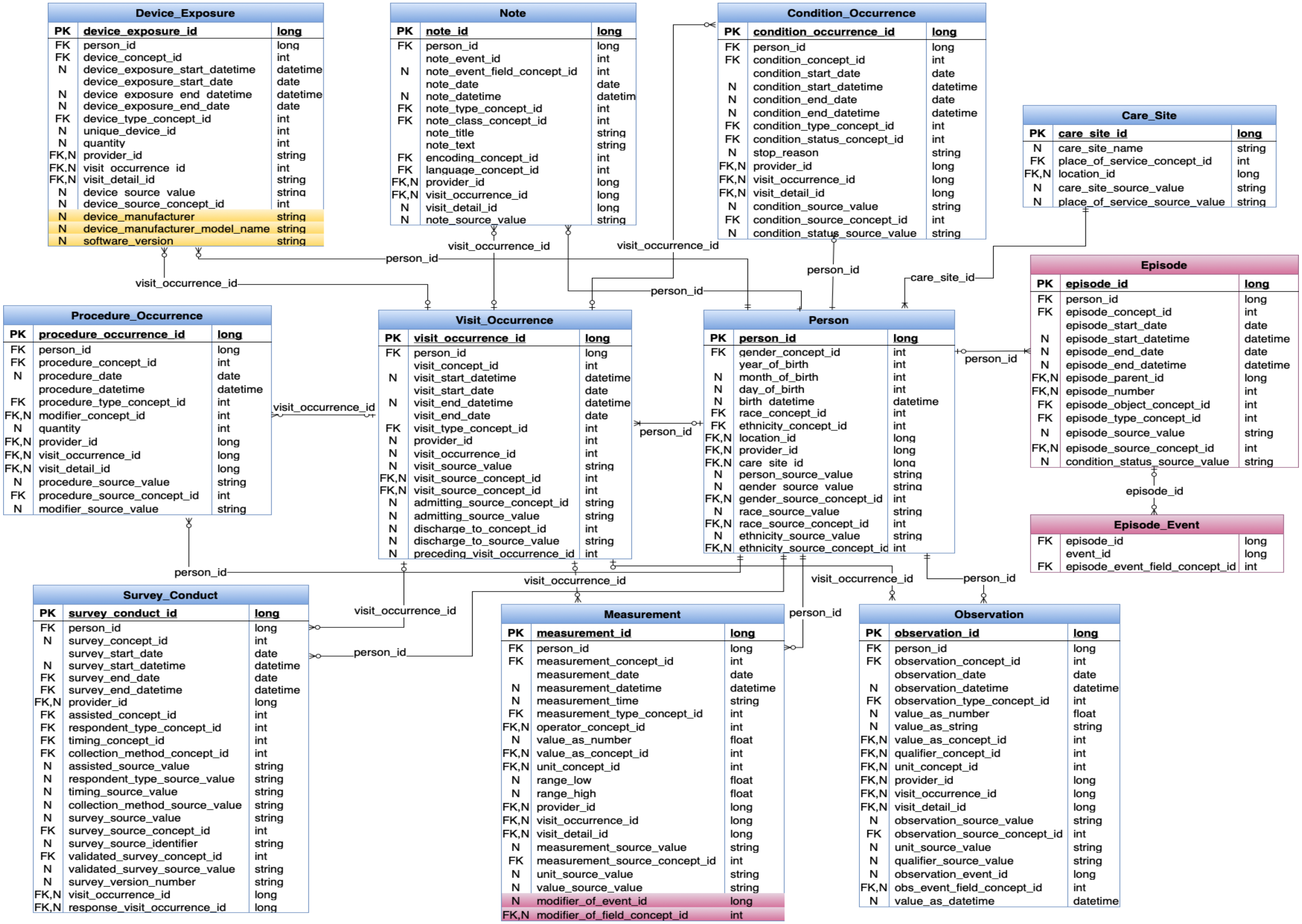
For all these clinical scenarios Imaging, Clinical Data and metadata will be uploaded to the ProstateNet of the ProCAncer-I platform using the ProCAncer-I upload tool which also acts as the communication gateway between the clinical site premises and the cloud platform.

Figure 11 shows the entity relationship diagram of the OMOP-CDM v6.0 as currently adopted in the metadata repository, along with its oncology extension. For space and clarity reasons, we haven't included the Vocabulary tables (i.e. Concept, Vocabulary, Domain, Concept_Class, Concept_Relationship, Relationship, Concept_Synonym, Concept_Ancestor) as well as the tables not instantiated by the data to be uploaded by the clinical sites (e.g. all health economics data tables, specimen, etc.). However, all the aforementioned vocabulary tables are incorporated into the metadata repository and instantiated with the use of the vocabularies downloaded by Athena[9].

Concerning the radiology image metadata accompanying each one of the use cases, we have designed and implemented an initial CDM-extension (based on the current R-CDM extension proposed by the OHDSI community) for storing all image related metadata required by the project, as depicted in Figure 12. In addition, we have designed an initial schema for storing the image curation information metadata as depicted in the same figure with yellow header tables.

To demonstrate how the data given by the clinical sites will be translated into the schemas described above, an example mapping of the raw data from the form "U1+2+3" to the OMOP-CDM is provided below. Figure 14 shows a screenshot of the "U1+2+3" e-CRF from the Upload Tool as it will be completed by the clinical experts and Figure 14 shows the mapping result to the OMOP-CDM tables with the corresponding concepts from the OMOP vocabulary. U1+2+3 eCRF describe cases with confirmed prostate cancer at biopsy and/or prostatectomy, with metastasis.

[9] https://athena.ohdsi.org/



Device_Exposure

PK	device_exposure_id	long
FK	person_id	long
FK	device_concept_id	int
N	device_exposure_start_datetime	datetime
	device_exposure_start_date	date
N	device_exposure_end_datetime	datetime
N	device_exposure_end_date	date
FK	device_type_concept_id	int
N	unique_device_id	int
N	quantity	int
FK,N	provider_id	string
FK,N	visit_occurrence_id	int
FK,N	visit_detail_id	string
N	device_source_value	string
N	device_source_concept_id	int
N	device_manufacturer	string
N	device_manufacturer_model_name	string
N	software_version	string

Note

PK	note_id	long
FK	person_id	long
	note_event_id	int
N	note_event_field_concept_id	int
	note_date	date
N	note_datetime	datetim
FK	note_type_concept_id	int
FK	note_class_concept_id	int
	note_title	string
	note_text	string
FK	encoding_concept_id	int
FK	language_concept_id	int
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
N	visit_detail_id	long
N	note_source_value	string

Condition_Occurrence

PK	condition_occurrence_id	long
FK	person_id	long
FK	condition_concept_id	int
	condition_start_date	date
N	condition_start_datetime	datetime
N	condition_end_date	date
N	condition_end_datetime	datetime
FK	condition_type_concept_id	int
FK	condition_status_concept_id	int
N	stop_reason	string
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
FK,N	visit_detail_id	long
N	condition_source_value	string
FK	condition_source_concept_id	int
N	condition_status_source_value	string

Care_Site

PK	care_site_id	long
N	care_site_name	string
FK	place_of_service_concept_id	int
FK,N	location_id	long
N	care_site_source_value	string
N	place_of_service_source_value	string

Episode

PK	episode_id	long
FK	person_id	long
FK	episode_concept_id	int
	episode_start_date	date
N	episode_start_datetime	datetime
N	episode_end_date	date
N	episode_end_datetime	datetime
FK,N	episode_parent_id	long
FK,N	episode_number	int
FK	episode_object_concept_id	int
FK	episode_type_concept_id	int
N	episode_source_value	string
FK,N	episode_source_concept_id	int
N	condition_status_source_value	string

Procedure_Occurrence

PK	procedure_occurrence_id	long
FK	person_id	long
FK	procedure_concept_id	int
N	procedure_date	date
	procedure_datetime	datetime
FK	procedure_type_concept_id	int
FK,N	modifier_concept_id	int
N	quantity	int
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
FK,N	visit_detail_id	long
N	procedure_source_value	string
FK	procedure_source_concept_id	int
N	modifier_source_value	string

Visit_Occurrence

PK	visit_occurrence_id	long
FK	person_id	long
	visit_concept_id	int
N	visit_start_datetime	datetime
	visit_start_date	date
N	visit_end_datetime	datetime
	visit_end_date	date
FK	visit_type_concept_id	int
N	provider_id	int
N	visit_occurrence_id	int
N	visit_source_value	string
FK,N	visit_source_concept_id	int
FK,N	visit_source_concept_id	int
N	admitting_source_concept_id	string
N	admitting_source_value	string
N	discharge_to_concept_id	int
N	discharge_to_source_value	string
N	preceding_visit_occurrence_id	int

Person

PK	person_id	long
FK	gender_concept_id	int
	year_of_birth	int
N	month_of_birth	int
N	day_of_birth	int
N	birth_datetime	datetime
FK	race_concept_id	int
FK	ethnicity_concept_id	int
FK,N	location_id	long
FK,N	provider_id	long
FK,N	care_site_id	long
N	person_source_value	string
N	gender_source_value	string
FK,N	gender_source_concept_id	int
N	race_source_value	string
FK,N	race_source_concept_id	int
N	ethnicity_source_value	string
FK,N	ethnicity_source_concept_id	int

Episode_Event

FK	episode_id	long
	event_id	long
FK	episode_event_field_concept_id	int

Survey_Conduct

PK	survey_conduct_id	long
FK	person_id	long
N	survey_concept_id	int
	survey_start_date	date
N	survey_start_datetime	datetime
FK	survey_end_date	date
FK	survey_end_datetime	datetime
FK,N	provider_id	long
FK	assisted_concept_id	int
FK	respondent_type_concept_id	int
FK	timing_concept_id	int
FK	collection_method_concept_id	int
N	assisted_source_value	string
N	respondent_type_source_value	string
N	timing_source_value	string
N	collection_method_source_value	string
N	survey_source_value	string
FK	survey_source_concept_id	int
N	survey_source_identifier	string
FK	validated_survey_concept_id	int
N	validated_survey_source_value	string
N	survey_version_number	string
FK,N	visit_occurrence_id	long
FK,N	response_visit_occurrence_id	long

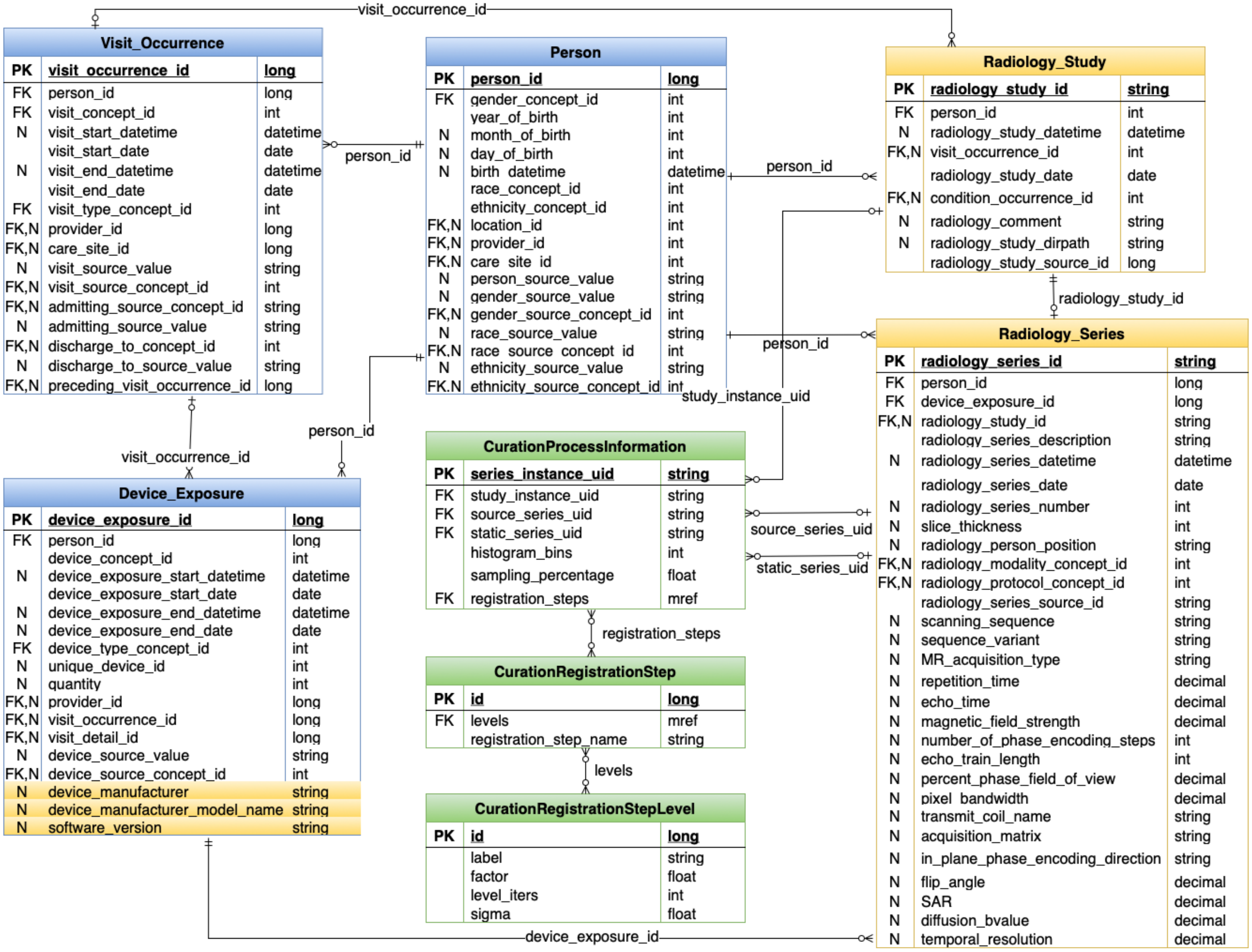
Measurement

PK	measurement_id	long
FK	person_id	long
FK	measurement_concept_id	int
	measurement_date	date
N	measurement_datetime	datetime
N	measurement_time	string
FK	measurement_type_concept_id	int
FK,N	operator_concept_id	int
N	value_as_number	float
FK,N	value_as_concept_id	int
FK,N	unit_concept_id	int
N	range_low	float
N	range_high	float
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
FK,N	visit_detail_id	long
N	measurement_source_value	string
FK	measurement_source_concept_id	int
N	unit_source_value	string
N	value_source_value	string
N	modifier_of_event_id	long
FK,N	modifier_of_field_concept_id	int

Observation

PK	observation_id	long
FK	person_id	long
FK	observation_concept_id	int
	observation_date	date
N	observation_datetime	datetime
FK	observation_type_concept_id	int
N	value_as_number	float
N	value_as_string	string
FK,N	value_as_concept_id	int
FK,N	qualifier_concept_id	int
FK,N	unit_concept_id	int
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
FK,N	visit_detail_id	long
N	observation_source_value	string
FK	observation_source_concept_id	int
N	unit_source_value	string
N	qualifier_source_value	string
N	observation_event_id	long
FK,N	obs_event_field_concept_id	int
N	value_as_datetime	datetime

Figure 11: Entity Relationship diagram of OMOP-CDM v6.0 and the oncology extension as adopted in the metadata repository.



Visit_Occurrence

PK	visit_occurrence_id	long
FK	person_id	long
FK	visit_concept_id	int
N	visit_start_datetime	datetime
	visit_start_date	date
N	visit_end_datetime	datetime
	visit_end_date	date
FK	visit_type_concept_id	int
FK,N	provider_id	long
FK,N	care_site_id	long
N	visit_source_value	string
FK,N	visit_source_concept_id	int
FK,N	admitting_source_concept_id	string
N	admitting_source_value	string
FK,N	discharge_to_concept_id	int
N	discharge_to_source_value	string
FK,N	preceding_visit_occurrence_id	long

Person

PK	person_id	long
FK	gender_concept_id	int
	year_of_birth	int
N	month_of_birth	int
N	day_of_birth	int
N	birth_datetime	datetime
	race_concept_id	int
	ethnicity_concept_id	int
FK,N	location_id	int
FK,N	provider_id	int
FK,N	care_site_id	int
N	person_source_value	string
N	gender_source_value	string
FK,N	gender_source_concept_id	int
N	race_source_value	string
FK,N	race_source_concept_id	int
N	ethnicity_source_value	string
FK.N	ethnicity_source_concept_id	int

Radiology_Study

PK	radiology_study_id	string
FK	person_id	int
N	radiology_study_datetime	datetime
FK,N	visit_occurrence_id	int
	radiology_study_date	date
FK,N	condition_occurrence_id	int
N	radiology_comment	string
N	radiology_study_dirpath	string
	radiology_study_source_id	long

Radiology_Series

PK	radiology_series_id	string
FK	person_id	long
FK	device_exposure_id	long
FK,N	radiology_study_id	string
	radiology_series_description	string
N	radiology_series_datetime	datetime
	radiology_series_date	date
N	radiology_series_number	int
N	slice_thickness	int
N	radiology_person_position	string
FK,N	radiology_modality_concept_id	int
FK,N	radiology_protocol_concept_id	int
	radiology_series_source_id	string
N	scanning_sequence	string
N	sequence_variant	string
N	MR_acquisition_type	string
N	repetition_time	decimal
N	echo_time	decimal
N	magnetic_field_strength	decimal
N	number_of_phase_encoding_steps	int
N	echo_train_length	int
N	percent_phase_field_of_view	decimal
N	pixel_bandwidth	decimal
N	transmit_coil_name	string
N	acquisition_matrix	string
N	in_plane_phase_encoding_direction	string
N	flip_angle	decimal
N	SAR	decimal
N	diffusion_bvalue	decimal
N	temporal_resolution	decimal

Device_Exposure

PK	device_exposure_id	long
FK	person_id	long
	device_concept_id	int
N	device_exposure_start_datetime	datetime
	device_exposure_start_date	date
N	device_exposure_end_datetime	datetime
N	device_exposure_end_date	date
FK	device_type_concept_id	int
N	unique_device_id	int
N	quantity	int
FK,N	provider_id	long
FK,N	visit_occurrence_id	long
FK,N	visit_detail_id	long
N	device_source_value	string
FK,N	device_source_concept_id	int
N	device_manufacturer	string
N	device_manufacturer_model_name	string
N	software_version	string

CurationProcessInformation

PK	series_instance_uid	string
FK	study_instance_uid	string
FK	source_series_uid	string
FK	static_series_uid	string
	histogram_bins	int
	sampling_percentage	float
FK	registration_steps	mref

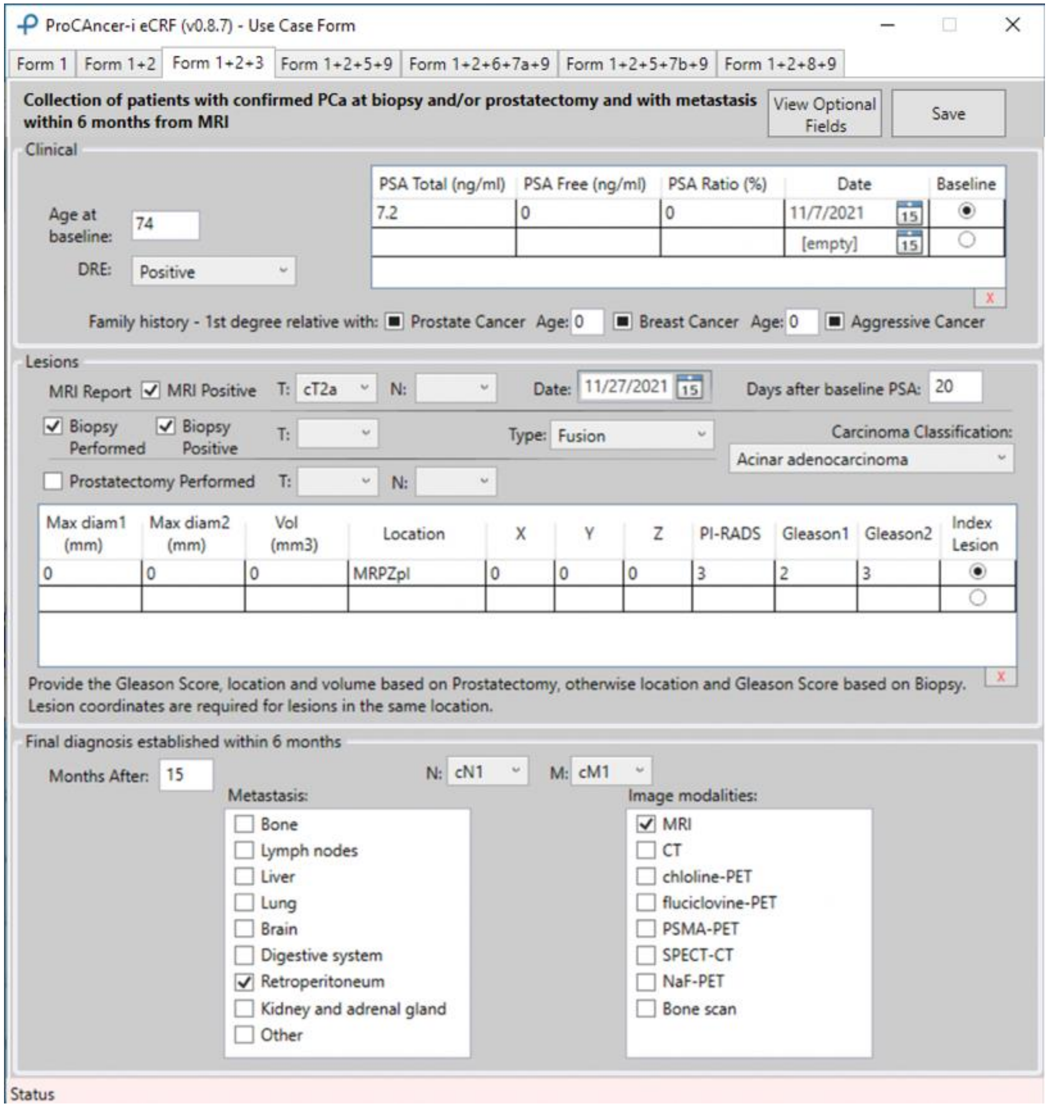
CurationRegistrationStep

PK	id	long
FK	levels	mref
	registration_step_name	string

CurationRegistrationStepLevel

PK	id	long
	label	string
	factor	float
	level_iters	int
	sigma	float

Figure 12: Entity Relationship diagram of CDM extension for Radiology and Curation Information



Collection of patients with confirmed PCa at biopsy and/or prostatectomy and with metastasis within 6 months from MRI

Clinical

Age at baseline: 74

DRE: Positive

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
7.2	0	0	11/7/2021	●
			[empty]	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report ☑ MRI Positive T: cT2a N: Date: 11/27/2021 Days after baseline PSA: 20

☑ Biopsy Performed ☑ Biopsy Positive T: Type: Fusion Carcinoma Classification: Acinar adenocarcinoma

☐ Prostatectomy Performed T: N:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
0	0	0	MRPZpl	0	0	0	3	2	3	●
										○

Provide the Gleason Score, location and volume based on Prostatectomy, otherwise location and Gleason Score based on Biopsy. Lesion coordinates are required for lesions in the same location.

Final diagnosis established within 6 months

Months After: 15 N: cN1 M: cM1

Metastasis:

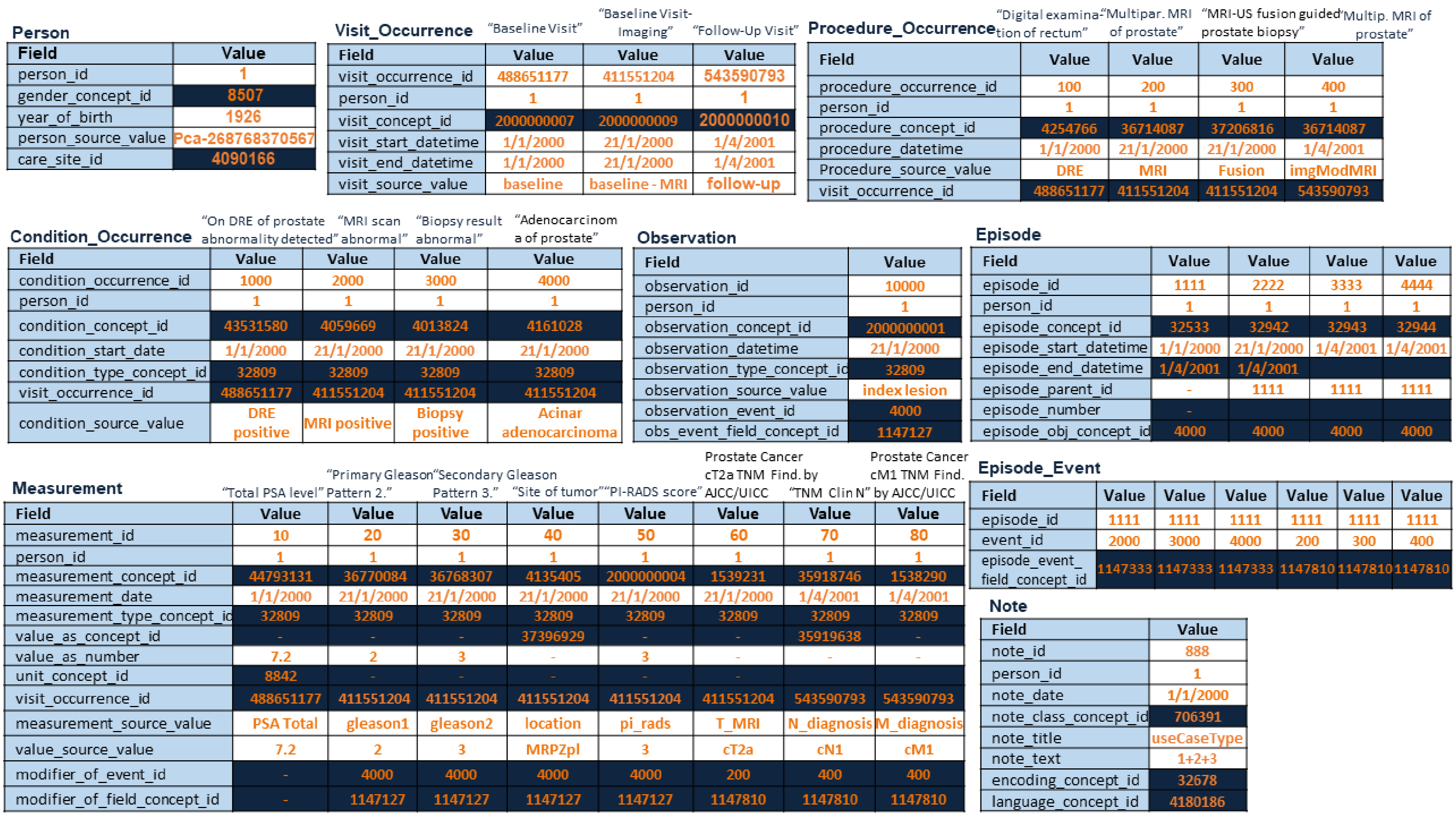
- ☐ Bone
- ☐ Lymph nodes
- ☐ Liver
- ☐ Lung
- ☐ Brain
- ☐ Digestive system
- ☑ Retroperitoneum
- ☐ Kidney and adrenal gland
- ☐ Other

Image modalities:

- ☑ MRI
- ☐ CT
- ☐ chloline-PET
- ☐ fluciclovine-PET
- ☐ PSMA-PET
- ☐ SPECT-CT
- ☐ NaF-PET
- ☐ Bone scan

Status

Figure 13: Screenshot of the eCRF Upload Tool for U1+2+3 form.



Person

Field	Value
person_id	1
gender_concept_id	8507
year_of_birth	1926
person_source_value	Pca-268768370567
care_site_id	4090166

Visit_Occurrence

	"Baseline Visit"	"Baseline Visit-Imaging"	"Follow-Up Visit"
Field	Value	Value	Value
visit_occurrence_id	488651177	411551204	543590793
person_id	1	1	1
visit_concept_id	2000000007	2000000009	2000000010
visit_start_datetime	1/1/2000	21/1/2000	1/4/2001
visit_end_datetime	1/1/2000	21/1/2000	1/4/2001
visit_source_value	baseline	baseline - MRI	follow-up

Procedure_Occurrence

	"Digital examina-tion of rectum"	"Multipar. MRI of prostate"	"MRI-US fusion guided prostate biopsy"	"Multip. MRI of prostate"
Field	Value	Value	Value	Value
procedure_occurrence_id	100	200	300	400
person_id	1	1	1	1
procedure_concept_id	4254766	36714087	37206816	36714087
procedure_datetime	1/1/2000	21/1/2000	21/1/2000	1/4/2001
Procedure_source_value	DRE	MRI	Fusion	imgModMRI
visit_occurrence_id	488651177	411551204	411551204	543590793

Condition_Occurrence

	"On DRE of prostate abnormality detected"	"MRI scan abnormal"	"Biopsy result abnormal"	"Adenocarcinoma of prostate"
Field	Value	Value	Value	Value
condition_occurrence_id	1000	2000	3000	4000
person_id	1	1	1	1
condition_concept_id	43531580	4059669	4013824	4161028
condition_start_date	1/1/2000	21/1/2000	21/1/2000	21/1/2000
condition_type_concept_id	32809	32809	32809	32809
visit_occurrence_id	488651177	411551204	411551204	411551204
condition_source_value	DRE positive	MRI positive	Biopsy positive	Acinar adenocarcinoma

Observation

Field	Value
observation_id	10000
person_id	1
observation_concept_id	2000000001
observation_datetime	21/1/2000
observation_type_concept_id	32809
observation_source_value	index lesion
observation_event_id	4000
obs_event_field_concept_id	1147127

Episode

Field	Value	Value	Value	Value
episode_id	1111	2222	3333	4444
person_id	1	1	1	1
episode_concept_id	32533	32942	32943	32944
episode_start_datetime	1/1/2000	21/1/2000	1/4/2001	1/4/2001
episode_end_datetime	1/4/2001	1/4/2001		
episode_parent_id	-	1111	1111	1111
episode_number	-			
episode_obj_concept_id	4000	4000	4000	4000

Measurement

	"Total PSA level"	"Primary Gleason Pattern 2."	"Secondary Gleason Pattern 3."	"Site of tumor"	"PI-RADS score"	Prostate Cancer cT2a TNM Find. by AJCC/UICC	"TNM Clin N"	Prostate Cancer cM1 TNM Find. by AJCC/UICC
Field	Value	Value	Value	Value	Value	Value	Value	Value
measurement_id	10	20	30	40	50	60	70	80
person_id	1	1	1	1	1	1	1	1
measurement_concept_id	44793131	36770084	36768307	4135405	2000000004	1539231	35918746	1538290
measurement_date	1/1/2000	21/1/2000	21/1/2000	21/1/2000	21/1/2000	21/1/2000	1/4/2001	1/4/2001
measurement_type_concept_id	32809	32809	32809	32809	32809	32809	32809	32809
value_as_concept_id	-	-	-	37396929	-	-	35919638	-
value_as_number	7.2	2	3	-	3	-	-	-
unit_concept_id	8842	-	-	-	-	-		
visit_occurrence_id	488651177	411551204	411551204	411551204	411551204	411551204	543590793	543590793
measurement_source_value	PSA Total	gleason1	gleason2	location	pi_rads	T_MRI	N_diagnosis	M_diagnosis
value_source_value	7.2	2	3	MRPZpl	3	cT2a	cN1	cM1
modifier_of_event_id	-	4000	4000	4000	4000	200	400	400
modifier_of_field_concept_id	-	1147127	1147127	1147127	1147127	1147810	1147810	1147810

Episode_Event

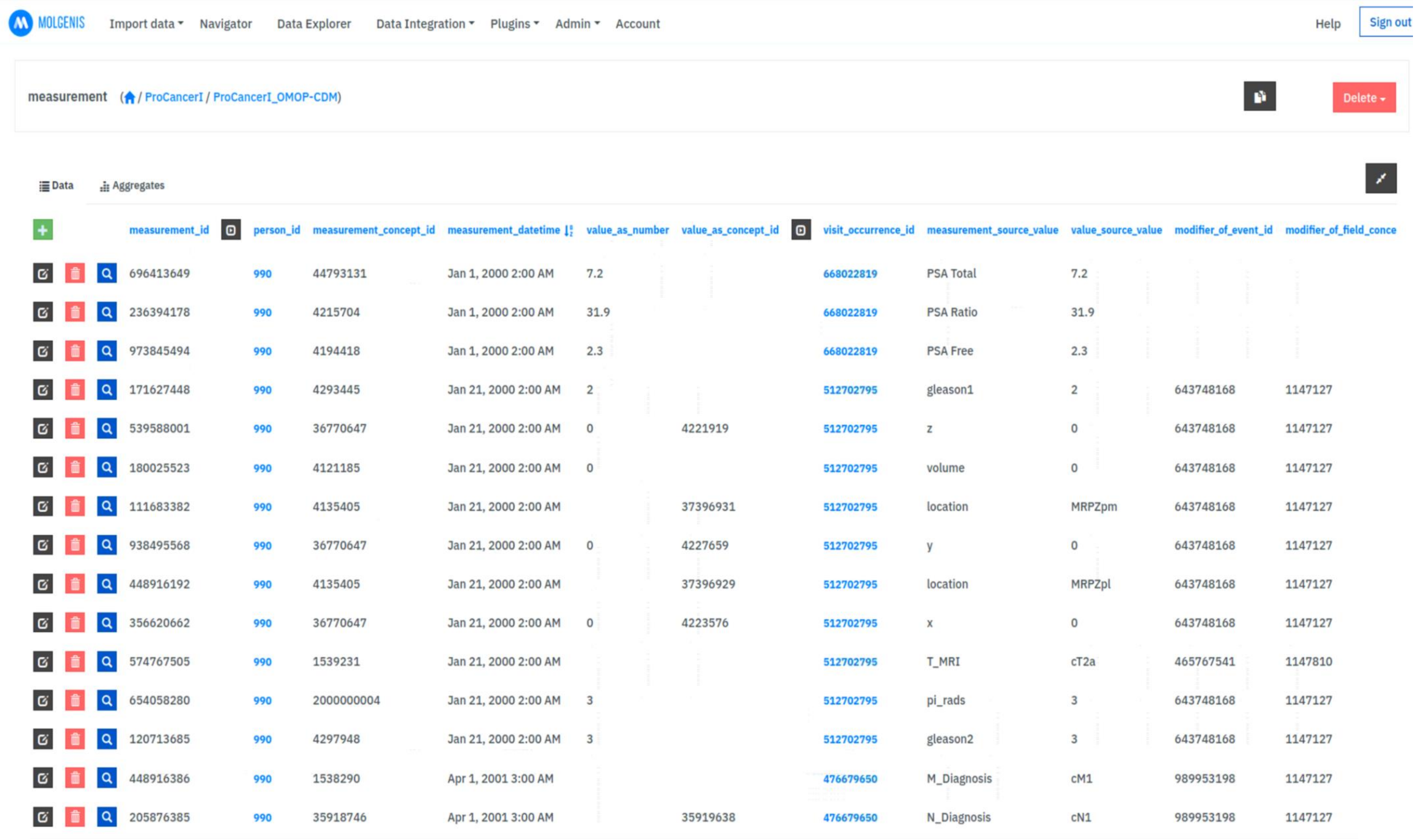
Field	Value	Value	Value	Value	Value	Value
episode_id	1111	1111	1111	1111	1111	1111
event_id	2000	3000	4000	200	300	400
episode_event_ field_concept_id	1147333	1147333	1147333	1147810	1147810	1147810

Note

Field	Value
note_id	888
person_id	1
note_date	1/1/2000
note_class_concept_id	706391
note_title	useCaseType
note_text	1+2+3
encoding_concept_id	32678
language_concept_id	4180186

Figure 14. Mapping of U1+2+3 form raw data into the OMOP-CDM

An example of the OMOP-CDM measurement metadata as stored in the Molgenis application is shown in Figure 15.



measurement (/ ProCancerI / ProCancerI_OMOP-CDM)

measurement_id	person_id	measurement_concept_id	measurement_datetime	value_as_number	value_as_concept_id	visit_occurrence_id	measurement_source_value	value_source_value	modifier_of_event_id	modifier_of_field_conce
696413649	990	44793131	Jan 1, 2000 2:00 AM	7.2		668022819	PSA Total	7.2		
236394178	990	4215704	Jan 1, 2000 2:00 AM	31.9		668022819	PSA Ratio	31.9		
973845494	990	4194418	Jan 1, 2000 2:00 AM	2.3		668022819	PSA Free	2.3		
171627448	990	4293445	Jan 21, 2000 2:00 AM	2		512702795	gleason1	2	643748168	1147127
539588001	990	36770647	Jan 21, 2000 2:00 AM	0	4221919	512702795	z	0	643748168	1147127
180025523	990	4121185	Jan 21, 2000 2:00 AM	0		512702795	volume	0	643748168	1147127
111683382	990	4135405	Jan 21, 2000 2:00 AM		37396931	512702795	location	MRPZpm	643748168	1147127
938495568	990	36770647	Jan 21, 2000 2:00 AM	0	4227659	512702795	y	0	643748168	1147127
448916192	990	4135405	Jan 21, 2000 2:00 AM		37396929	512702795	location	MRPZpl	643748168	1147127
356620662	990	36770647	Jan 21, 2000 2:00 AM	0	4223576	512702795	x	0	643748168	1147127
574767505	990	1539231	Jan 21, 2000 2:00 AM			512702795	T_MRI	cT2a	465767541	1147810
654058280	990	2000000004	Jan 21, 2000 2:00 AM	3		512702795	pi_rads	3	643748168	1147127
120713685	990	4297948	Jan 21, 2000 2:00 AM	3		512702795	gleason2	3	643748168	1147127
448916386	990	1538290	Apr 1, 2001 3:00 AM			476679650	M_Diagnosis	cM1	989953198	1147127
205876385	990	35918746	Apr 1, 2001 3:00 AM		35919638	476679650	N_Diagnosis	cN1	989953198	1147127

Figure 15. OMOP-CDM Measurement table instances

Conventions:

We adopted the following conventions as concerning the dates used in the model, the primary diagnosis stored when this is not captured in the data, as well as a set of "visit" concepts for achieving better grouping of the data.

Dates:

- We consider 01/01/2000 as the baseline date. All dates are calculated based on this date.
- For the procedures that dates are missing (e.g., biopsy, prostatectomy), we use the same date as the date the MRI was performed. If procedures refer to follow-ups, then we use the follow-up date of the MRI.

Condition Occurrence:

- In some use cases where the primary diagnosis is not captured in the data itself but there is a positive biopsy, a condition occurrence record "Primary malignant neoplasm of prostate" (200962) is stored. Otherwise, a "lesion of prostate" (4115735) is stored for every lesion that exists in the data.

Study Visits:

We have defined multiple custom concepts to capture the full granularity of study visits and achieve grouping of the data. For example:

- **Baseline Visit (2000000007):** the PSA visit marked as the baseline.
- **PSA Visit (2000000008)**: any of the additional PSA visits following/preceding the baseline one.
- **Baseline Visit - Imaging (2000000009)**: the MRI visit, for which there is a date specified.
- **Follow-up Visit (2000000010):** Follow-up visit, when there is no information about the dates that exams/procedures were performed.
- **Follow-Up Visit - Baseline (2000000011):** For the use cases that there are PSA exams as follow-ups, this visit describes the baseline one.
- **Follow-Up Visit - PSA (2000000012):** any of the additional PSA visits following/preceding the follow-up baseline one.
- **Follow-Up Visit - Biopsy (2000000013):** the Biopsy visit as a follow up, for which there is a date specified.
- **Follow-Up Visit - Imaging (2000000014):** the MRI visit as a follow-up, for which there is a date specified

Metadata catalogue APIs

The Metadata catalogue of ProstateNet is accessible through a web-based ("RESTful") API, provided at:

https://metadata.procancer-i.eu/capi

Currently, it supports only “Store” endpoints for storing each use case’s clinical and imaging metadata information as well as curation information metadata. Each request should be authenticated through the ProCAncer-I’s Identity Provider (see section 5.2) and convey an “access token” (JSON Web Token) with specific claims, such as the audience (“aud”), which should be “ProstateNet”, and the client’s identification.

The following are the required HTTP headers and request parameters:

- `Content-Type: multipart/form-data`
- `Authorization: Bearer` *`{access token}`*
- `data_file`: the JSON file containing the information about the respective UC (for the ../usecases endpoint) or the curation related metadata information (for the ../curationMeta endpoint)

Table 1. Metadata Catalogue API Endpoints

Method	Path	Description
POST	`../usecases`	Store all information related to the respective use case.
POST	../curationMeta	Store curation related metadata for the co-registration and motion correction procedures.

Table 2. Metadata Catalogue Store Response Status Codes

Code	Description
200 (OK)	All data for the uploaded use case has been stored.
202 (Accepted)	Some of the information in the request has been stored but others have failed.
400 (Bad Request)	The request was badly formatted. For example, the provided JSON file is invalid.
401 (Unauthorized)	The client is not authenticated.
503 (Service Unavailable)	The connection to the metadata catalogue could not be established.

5. Services & Tools

This chapter details the implementation of all services required as part of the alpha version of the platform, both internal and user-oriented, that are part of the ProCAncer-I Platform. It describes the platform services for authentication and authorization backends and middleware. It also describes user-oriented tools for data anonymization, upload, sharing, curation, and annotation. It also provides details of the various tools' APIs, the Image Repository's DICOMweb-compliant API, including response payloads and status codes, and the technologies used to develop each service.

5.1 DICOMweb API

The ProCAncer-I Image Repository API is based on the open source DICOM server **Medical Imaging Server for DICOM** (https://github.com/microsoft/dicom-server) that allows standards-based communication with any DICOMweb™ enabled systems and supports a subset of the DICOMweb Standard (https://www.dicomstandard.org/dicomweb):

- Store (STOW-RS)
- Retrieve (WADO-RS)
- Search (QIDO-RS)

Delete and Retrieve of metadata is not supported.

The API is provided at: https://procancer-i.3dnetmedical.com:63801.

RESTful services

Store (STOW-RS)

This transaction uses the POST method to Store representations of Studies, Series, and Instances contained in the request payload.

Table 3. API Gateway STOW-RS Endpoints

Method	Path	Description
POST	…/studies	Store instances.
POST	…/studies/*{study}*	Store instances for a specific study.

Parameter study corresponds to the DICOM attribute `StudyInstanceUID`. If specified, any instance that does not belong to the provided study will be rejected with 43265 warning code.

The following Accept header(s) for the response are supported:

- `application/dicom+json`

The following Content-Type header(s) are supported:

- `multipart/related; type="application/dicom"`
- `application/dicom`

Note: The Server will not coerce or replace attributes that conflict with existing data. All data will be stored as provided.

The following DICOM elements are required to be present in every DICOM file attempting to be stored:

- `StudyInstanceUID`
- `SeriesInstanceUID`
- `SOPInstanceUID`
- `SOPClassUID`
- `PatientID`

Note: **All identifiers must be between 1 and 64 characters long**, and only contain alphanumeric characters or the following special characters: '.', '-'.

Each file stored must have a unique combination of `StudyInstanceUID`, `SeriesInstanceUID` and `SOPInstanceUID`. The warning code 45070 will be returned if a file with the same identifiers already exists.

DICOM File Size Limit: there is a size limit of 2GB for a DICOM file by default.

Table 4. API Gateway STOW-RS Response Status Codes

Code	Description
200 (OK)	All the SOP instances in the request have been stored.
202 (Accepted)	Some instances in the request have been stored but others have failed.
204 (No Content)	No content was provided in the store transaction request.
400 (Bad Request)	The request was badly formatted. For example, the provided study instance identifier did not conform to the expected UID format.
401 (Unauthorized)	The client is not authenticated.
406 (Not Acceptable)	The specified Accept header is not supported.

Code	Description
409 (Conflict)	None of the instances in the store transaction request have been stored.
415 (Unsupported Media Type)	The provided Content-Type is not supported.
503 (Service Unavailable)	The service is unavailable or busy. Please try again later.

The response payload will populate a DICOM dataset with the following elements:

Table 5. API Gateway STOW-RS Response Payload

Tag	Name	Description
(0008, 1190)	`RetrieveURL`	The Retrieve URL of the study if the `StudyInstanceUID` was provided in the store request and at least one instance is successfully stored.
(0008, 1198)	`FailedSOPSequence`	The sequence of instances that failed to store.
(0008, 1199)	`ReferencedSOPSequence`	The sequence of stored instances.

Each dataset in the `FailedSOPSequence` will have the following elements (if the DICOM file attempting to be stored could be read):

Table 6. API Gateway STOW-RS Response Dataset Elements

Tag	Name	Description
(0008, 1150)	`ReferencedSOPClassUID`	The SOP class unique identifier of the instance that failed to store.
(0008, 1155)	`ReferencedSOPInstanceUID`	The SOP instance unique identifier of the instance that failed to store.
(0008, 1197)	`FailureReason`	The reason code why this instance failed to store.

Each dataset in the `ReferencedSOPSequence` will have the following elements:

Table 7. API Gateway STOW-RS Response Dataset Elements

Tag	Name	Description
(0008, 1150)	`ReferencedSOPClassUID`	The SOP class unique identifier of the instance that failed to store.
(0008, 1155)	`ReferencedSOPInstanceUID`	The SOP instance unique identifier of the instance that failed to store.
(0008, 1190)	`RetrieveURL`	The retrieve URL of this instance on the DICOM server.

An example response with Accept header application/dicom+json:

```
{
    "00081190": {
      "vr": "UR",
      "Value": [
        "http://localhost/studies/d09e8215-e1e1-4c7a-8496-b4f6641ed232"
      ]
    },
    "00081198": {
      "vr": "SQ",
      "Value": [
        {
          "00081150": {
            "vr": "UI",
            "Value": [
              "cd70f89a-05bc-4dab-b6b8-1f3d2fcafeec"
            ]
          },
          "00081155": {
            "vr": "UI",
            "Value": [
              "22c35d16-11ce-43fa-8f86-90ceed6cf4e7"
            ]
          },
          "00081197": {
            "vr": "US",
            "Value": [
              43265
            ]
          }
        }
      ]
    },
    "00081199": {
      "vr": "SQ",
      "Value": [
        {
          "00081150": {
            "vr": "UI",
            "Value": [
              "d246deb5-18c8-4336-a591-aeb6f8596664"
            ]
          },
          "00081155": {
            "vr": "UI",
            "Value": [
              "4a858cbb-a71f-4c01-b9b5-85f88b031365"
            ]
          },
          "00081190": {
            "vr": "UR",
            "Value": [
              "http://localhost/studies/d09e8215-e1e1-4c7a-8496-
b4f6641ed232/series/8c4915f5-cc54-4e50-aa1f-9b06f6e58485/instances/4a858cbb-
a71f-4c01-b9b5-85f88b031365"
            ]
          }
        }
      ]
    }
  }
```

Table 8. *API Gateway STOW-RS Failure Reason Codes*

Code	Description
272	The store transaction did not store the instance because of a general failure in processing the operation.
43264	The DICOM instance failed the validation.
43265	The provided instance `StudyInstanceUID` did not match the specified `StudyInstanceUID` in the store request.
45070	A DICOM instance with the same `StudyInstanceUID`, `SeriesInstanceUID` and `SopInstanceUID` has already been stored. If you wish to update the contents, delete this instance first.
45071	A DICOM instance is being created by another process, or the previous attempt to create has failed and the cleanup process has not had the chance to clean up yet. Please delete the instance first before attempting to create it again.

Retrieve (WADO-RS)

This Retrieve Transaction offers support for retrieving stored studies, series, instances and frames by reference.

Table 9. API Gateway WADO-RS Endpoints

Method	Path	Description
GET	`…/studies/{study}`	Retrieves all instances within a study.
GET	`…/studies/{study}/series/{series}`	Retrieves all instances within a series.
GET	`…/studies/{study}/series/{series}/instances/{instance}`	Retrieves a single instance.
GET	`…/studies/{study}/series/{series}/instances/{instance}/frames/{frames}`	Retrieves one or many frames from a single instance. To specify more than one frame, a comma separate each frame to return, e.g. `/studies/1/series/2/instance/3/frames/4,5,6`

Retrieve instances within Study or Series

The following Accept header(s) are supported for retrieving instances within a study or a series:

- `multipart/related; type="application/dicom"; transfer-syntax=*`
- `multipart/related; type="application/dicom";` (when transfer-syntax is not specified, `1.2.840.10008.1.2.1` is used as default)
- `multipart/related; type="application/dicom"; transfer-syntax=1.2.840.10008.1.2.1`
- `multipart/related; type="application/dicom"; transfer-syntax=1.2.840.10008.1.2.4.90`

Retrieve an Instance

The following Accept header(s) are supported for retrieving a specific instance:

- `application/dicom; transfer-syntax=*`
- `multipart/related; type="application/dicom"; transfer-syntax=*`
- `application/dicom;` (when transfer-syntax is not specified, `1.2.840.10008.1.2.1` is used as default)
- `multipart/related; type="application/dicom"` (when transfer-syntax is not specified, `1.2.840.10008.1.2.1` is used as default)
- `application/dicom; transfer-syntax=1.2.840.10008.1.2.1`
- `multipart/related; type="application/dicom"; transfer-syntax=1.2.840.10008.1.2.1`
- `application/dicom; transfer-syntax=1.2.840.10008.1.2.4.90`
- `multipart/related; type="application/dicom"; transfer-syntax=1.2.840.10008.1.2.4.90`

Retrieve Frames

The following Accept headers are supported for retrieving frames:

- `multipart/related; type="application/octet-stream"; transfer-syntax=*`
- `multipart/related; type="application/octet-stream";` (when transfer-syntax is not specified, `1.2.840.10008.1.2.1` is used as default)
- `multipart/related; type="application/octet-stream"; transfer-syntax=1.2.840.10008.1.2.1`
- `multipart/related; type="image/jp2";` (when transfer-syntax is not specified, `1.2.840.10008.1.2.4.90` is used as default)
- `multipart/related; type="image/jp2";transfer-syntax=1.2.840.10008.1.2.4.90`

Retrieve Transfer Syntax

When the requested transfer syntax is different from the original file, the original file is transcoded to the requested transfer syntax. The original file needs to be one of below formats for transcoding to succeed, otherwise transcoding may fail:

- 1.2.840.10008.1.2 (Little Endian Implicit)
- 1.2.840.10008.1.2.1 (Little Endian Explicit)
- 1.2.840.10008.1.2.2 (Explicit VR Big Endian)
- 1.2.840.10008.1.2.4.50 (JPEG Baseline Process 1)
- 1.2.840.10008.1.2.4.57 (JPEG Lossless)
- 1.2.840.10008.1.2.4.70 (JPEG Lossless Selection Value 1)
- 1.2.840.10008.1.2.4.90 (JPEG 2000 Lossless Only)
- 1.2.840.10008.1.2.4.91 (JPEG 2000)
- 1.2.840.10008.1.2.5 (RLE Lossless)

An unsupported transfer-syntax will result in 406 Not Acceptable.

Table 10. API Gateway WADO-RS Response Status Codes

Code	Description
200 (OK)	All requested data has been retrieved.
304 (Not Modified)	The requested data has not modified since the last request. Content is not added to the response body in such case. Please see Retrieve Metadata Cache Validation (for Study, Series, or Instance) for more information.
400 (Bad Request)	The request was badly formatted. For example, the provided study instance identifier did not conform the expected UID format or the requested transfer-syntax encoding is not supported.

401 (Unauthorized)	The client is not authenticated.
404 (Not Found)	The specified DICOM resource could not be found.
406 (Not Acceptable)	The specified Accept header is not supported.
503 (Service Unavailable)	The service is unavailable or busy. Please try again later.

Search (QIDO-RS)

Queries based on ID for DICOM Objects (QIDO) enable you to search for studies, series and instances by attributes.

Table 11. API Gateway QIDO-RS Endpoints

	Method	Path	Description
Search for Studies			
	GET	…/studies?…	Search for studies
Search for Series			
	GET	…/series?…	Search for series
	GET	…/studies/*{study}*/series?…	Search for series in a study
Search for Instances			
	GET	…/instances?…	Search for instances
	GET	…/studies/{study}/instances?…	Search for instances in a study
	GET	…/studies/*{study}*/series/*{series}*/instances? …	Search for instances in a series

The following Accept header(s) are supported for searching:

- application/dicom+json

The following parameters for each query are supported:

Table 12. API Gateway QIDO-RS Supported Search Parameters

Key	Support Value(s)	Allowed Count	Description
{attributeID}=	{value}	0...N	Search for attribute/value matching in query.

Key	Support Value(s)	Allowed Count	Description
includefield=	{attributeID} all	0...N	The additional attributes to return in the response. Both, public and private tags are supported. When all is provided, please see Search Response for more information about which attributes will be returned for each query type. If a mixture of {attributeID} and 'all' is provided, the server will default to using 'all'.
limit=	{value}	0..1	Integer value to limit the number of values returned in the response. Value can be between the range 1 >= x <= 200. Defaulted to 100.
offset=	{value}	0..1	Skip {value} results. If an offset is provided larger than the number of search query results, a 204 (no content) response will be returned.
fuzzymatching=	true | false	0..1	If true fuzzy matching is applied to PatientName attribute. It will do a prefix word match of any name part inside PatientName value. For example, if PatientName is "John^Doe", then "joh", "do", "jo do", "Doe" and "John Doe" will all match. However "ohn" will not match.

We support searching on the attributes and search type listed in Table 13.

Table 13. API Gateway QIDO-RS Searchable Attributes

Attribute Keyword	Study	Series	Instance
StudyInstanceUID	X	X	X
PatientName	X	X	X
PatientID	X	X	X
PatientBirthDate	X	X	X
AccessionNumber	X	X	X
ReferringPhysicianName	X	X	X
StudyDate	X	X	X
StudyDescription	X	X	X

Attribute Keyword	Study	Series	Instance
`SeriesInstanceUID`		X	X
`Modality`		X	X
`PerformedProcedureStepStartDate`		X	X
`ManufacturerModelName`		X	X
`SOPInstanceUID`			X

We support the matching types shown below.

Table 14. API Gateway QIDO-RS Search Matching

Search Type	Supported Attribute	Example
Range Query	StudyDate, PatientBirthDate	{attributeID}={value1}-{value2}. For date/ time values, we supported an inclusive range on the tag. This will be mapped to attributeID >= {value1} AND attributeID <= {value2}.
Exact Match	All supported attributes	{attributeID}={value1}
Fuzzy Match	PatientName, ReferringPhysicianName	Matches any component of the name which starts with the value.

Attribute ID

Tags can be encoded in a number of different ways for the query parameter. We have partially implemented the standard as defined in PS3.18 6.7.1.1.1. The following encodings for a tag are supported.

Table 15. API Gateway QIDO-RS Tag Encodings

Value	Example
{group}{element}	`0020000D`
{dicomKeyword}	`StudyInstanceUID`

Example query searching for instances:

```
…/instances?Modality=CT&00280011=512&includefield=00280010&limit=5&offset=0
```

Search Response

The response will be an array of DICOM datasets. Depending on the resource, by *default* the following attributes are returned:

Default Study tags

Table 16. API Gateway QIDO-RS Response Datasets' Attributes

Tag	Attribute Name
(0008, 0005)	`SpecificCharacterSet`
(0008, 0020)	`StudyDate`
(0008, 0030)	`StudyTime`
(0008, 0050)	`AccessionNumber`
(0008, 0056)	`InstanceAvailability`
(0009, 0090)	`ReferringPhysicianName`
(0008, 0201)	`TimezoneOffsetFromUTC`
(0010, 0010)	`PatientName`
(0010, 0020)	`PatientID`
(0010, 0030)	`PatientBirthDate`
(0010, 0040)	`PatientSex`
(0020, 0010)	`StudyID`
(0020, 000D)	`StudyInstanceUID`

Default Series tags

Table 17. API Gateway QIDO-RS Response Datasets' Attributes

Tag	Attribute Name
(0008, 0005)	`SpecificCharacterSet`
(0008, 0060)	`Modality`
(0008, 0201)	`TimezoneOffsetFromUTC`
(0008, 103E)	`SeriesDescription`
(0020, 000E)	`SeriesInstanceUID`
(0040, 0244)	`PerformedProcedureStepStartDate`
(0040, 0245)	`PerformedProcedureStepStartTime`
(0040, 0275)	`RequestAttributesSequence`

Default Instance tags

Table 18. API Gateway QIDO-RS Response Datasets' Attributes

Tag	Attribute Name
(0008, 0005)	`SpecificCharacterSet`
(0008, 0016)	`SOPClassUID`
(0008, 0018)	`SOPInstanceUID`
(0008, 0056)	`InstanceAvailability`
(0008, 0201)	`TimezoneOffsetFromUTC`
(0020, 0013)	`InstanceNumber`
(0028, 0010)	`Rows`
(0028, 0011)	`Columns`
(0028, 0100)	`BitsAllocated`
(0028, 0008)	`NumberOfFrames`

If includefield=all, below attributes are included along with default attributes. Along with default attributes, this is the full list of attributes supported at each resource level.

Additional Study tags

Table 19. API Gateway QIDO-RS Response Datasets' Attributes

Tag	Attribute Name
(0008, 1030)	`StudyDescription`
(0008, 0063)	`AnatomicRegionsInStudyCodeSequence`
(0008, 1032)	`ProcedureCodeSequence`
(0008, 1060)	`NameOfPhysiciansReadingStudy`
(0008, 1080)	`AdmittingDiagnosesDescription`
(0008, 1110)	`ReferencedStudySequence`
(0010, 1010)	`PatientAge`
(0010, 1020)	`PatientSize`
(0010, 1030)	`PatientWeight`
(0010, 2180)	`Occupation`

Tag	Attribute Name
(0010, 21B0)	AdditionalPatientHistory

Additional Series tags

Table 20. *API Gateway QIDO-RS Response Datasets' Attributes*

Tag	Attribute Name
(0020, 0011)	SeriesNumber
(0020, 0060)	Laterality
(0008, 0021)	SeriesDate
(0008, 0031)	SeriesTime

Along with those below attributes are returned:

- All the match query parameters and UIDs in the resource url.
- IncludeField attributes supported at that resource level.
- If the target resource is All Series, then Study level attributes are also returned.
- If the target resource is All Instances, then Study and Series level attributes are also returned.
- If the target resource is Study's Instances, then Series level attributes are also returned.

Search Response Codes

The query API will return one of the following status codes in the response:

Table 21. *API Gateway QIDO-RS Response Status Codes*

Code	Description
200 (OK)	The response payload contains all the matching resource.
204 (No Content)	The search completed successfully but returned no results.
400 (Bad Request)	The server was unable to perform the query because the query component was invalid. Response body contains details of the failure.
401 (Unauthorized)	The client is not authenticated.
503 (Service Unavailable)	The service is unavailable or busy. Please try again later.

Additional Notes

- Querying using the `TimezoneOffsetFromUTC` (00080201) is not supported.
- The query API will not return 413 (request entity too large). If the requested query response limit is outside of the acceptable range, a bad request will be returned. Anything requested within the acceptable range, will be resolved.
- When target resource is Study/Series there is a potential for inconsistent study/series level metadata across multiple instances. For example, two instances could have different `patientName`. In this case latest will win and you can search only on the latest data.
- Paged results are optimized to return matched *newest* instance first, this may result in duplicate records in subsequent pages if newer data matching the query was added.
- Matching is case in-sensitive and accent in-sensitive for PN VR types.
- Matching is case in-sensitive and accent sensitive for other string VR types.

5.2 User Authentication and Authorization

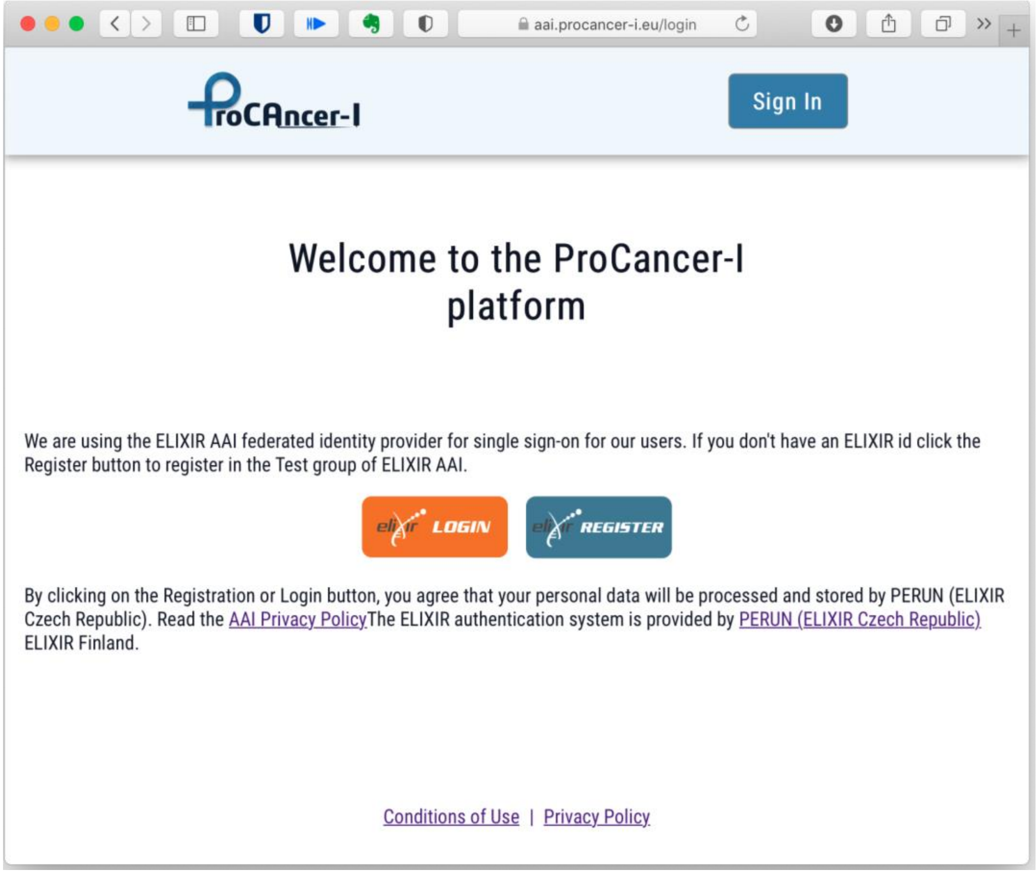
The ProCAncer-I technical platform consists of multiple interacting software components, and it is planned to also be accessed by a diverse set of users (e.g., clinicians, radiologists, modelers, etc.). In order to balance user-friendliness in accessing data and other platform services and the security requirements for “human to machine” and “machine to machine” communication, the following design decisions were made:

- Use authentication is implemented through a “single sign-on” scheme, i.e., the users do not need to re-authenticate when they access different services of the platform. Moreover, users can use their existing credentials provided by their home organizations in order to log into the platform with minimal effort
- Every service-to-service communication and API interaction is accompanied by secure (signed) tokens in order to authenticate the requesting service and make authorization decisions based on the principal user that triggered this interaction.
- Authorization decisions are made using the contents of the tokens and the users’ “claims” encoded therein, but also context information such as the requesting service, the type of data or functionality requested, etc. Since the rules for such authorization decisions can be complex and relevant to many different parts of the platform, they are managed in a centralized manner through a central authorization decision service. The access policies are enforced by the different services of the platform after contacting the central policy decision point.

In the following paragraphs we describe the main software components comprising the authentication and authorization infrastructure of the platform.

5.2.1 Identity Provider

The ProCAncer-I “Identity Provider” is the authentication and authorization proxy for the users and platform services. It is linked with the Authentication and Authorization Infrastructure (AAI) of ELIXIR and therefore it is supported by a federation of identity providers of research institutions and companies around Europe. The ELIXIR AAI[10] provides services for data providers and authorities for researcher identification, authentication and authorisation. The ELIXIR AAI assigns each end user a permanent unique identifier (ELIXIR ID), which allows them to use the same credentials (e.g., home university username and password) to login to services that belong to ELIXIR. The data access and resource allocations are stored as part of the identity. The ELIXIR AAI allows authentication against services associated with ELIXIR research infrastructure, and in the long term, it can be used with other research infrastructures and e-infrastructures. There is also an ongoing activity to migrate the ELIXIR AAI into the LifeScience AAI[11] which will provide similar service across the wider life sciences research community in Europe After the initial registration, the user selects its (ELIXIR integrated) identity provider and is forwarded to her/his "home organization" where s/he is authenticated with their "usual" authentication means (Figure 17). Single Sign-On is therefore supported since the users do not need to create additional usernames and passwords in order to login into the ProCAncer-I platform.



Welcome to the ProCancer-I platform

We are using the ELIXIR AAI federated identity provider for single sign-on for our users. If you don't have an ELIXIR id click the Register button to register in the Test group of ELIXIR AAI.

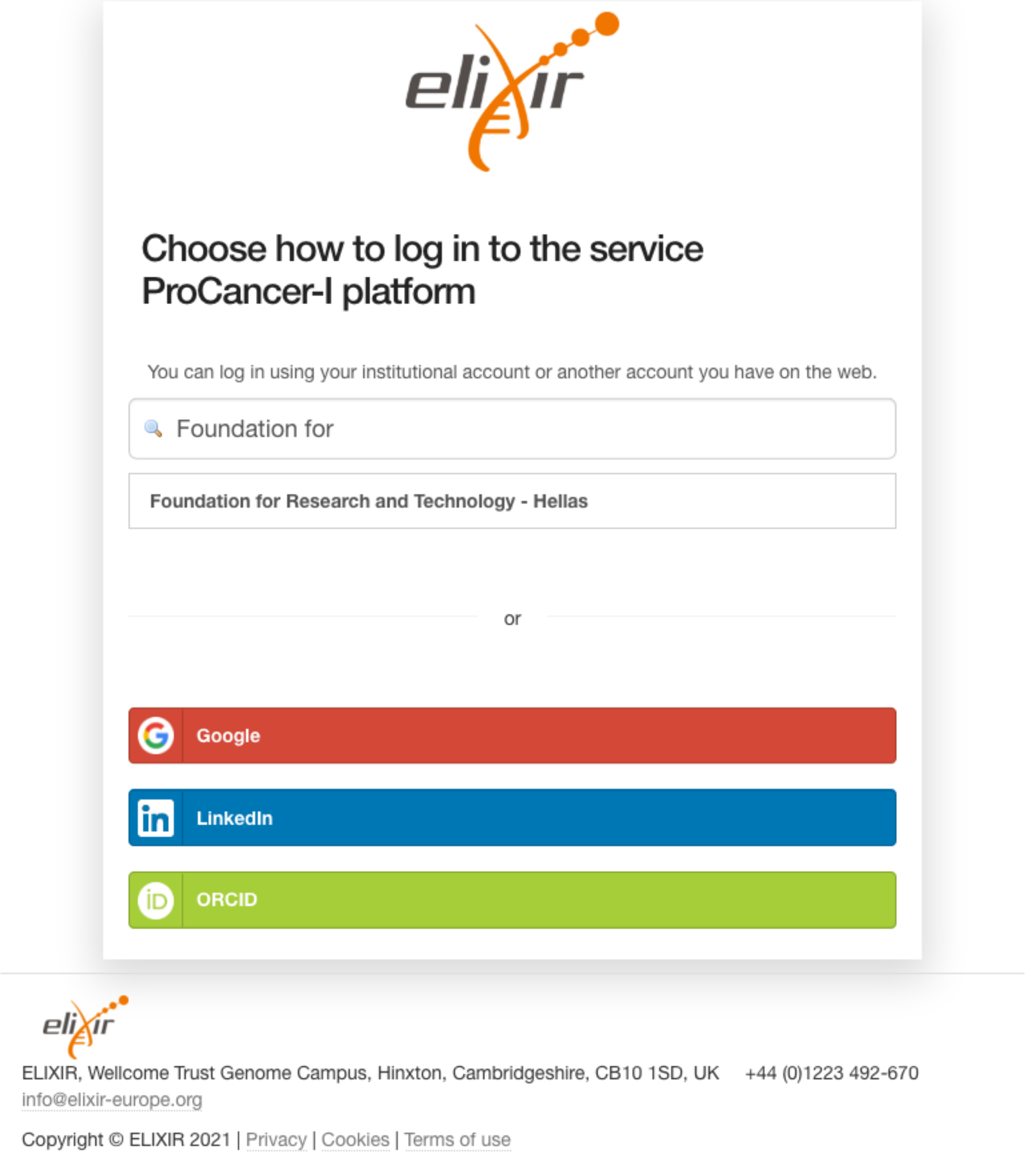
By clicking on the Registration or Login button, you agree that your personal data will be processed and stored by PERUN (ELIXIR Czech Republic). Read the AAI Privacy PolicyThe ELIXIR authentication system is provided by PERUN (ELIXIR Czech Republic) ELIXIR Finland.

Conditions of Use | Privacy Policy

Figure 16. ProstateNet Login page

[10] Linden M, Procházka M, Lappalainen I et al. Common ELIXIR Service for Researcher Authentication and Authorisation. F1000Research 2018, 7(ELIXIR):1199 https://doi.org/10.12688/f1000research.15161.1

[11] Linden, M., Bucik, D., Gormanns, et al. Access and User Management System for Life Science -- the Blueprint EOSC-Life Deliverable D5.1. https://doi.org/10.5281/zenodo.3386307



Choose how to log in to the service
ProCancer-I platform

You can log in using your institutional account or another account you have on the web.

Foundation for

Foundation for Research and Technology - Hellas

ELIXIR, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK +44 (0)1223 492-670
info@elixir-europe.org

Copyright © ELIXIR 2021 | Privacy | Cookies | Terms of use

Figure 17. Authentication through the ELIXIR Federated AAI and its integrated organizations

The ProCAncer-I *Identity Provider* operates as a client to the ELIXIR AAI, but for the rest of the platform it is the central authentication point using the OAuth 2.0 and OpenID Connect standards. Following OpenID Connect, a popular industry standard (used by Google, Facebook, and other big companies), it issues JSON Web Tokens (JWT) that provide “claims” (such as the users’ names, organizations, etc) to the requested services about the authenticated users. These tokens are signed by the Identity Provider using its private key and then they can be passed (as “Bearer” tokens) to other services that check their validity using the Id Provider’s public key. In terms of the standards supported the following list is not complete but nevertheless indicative:

- RFC 6749 “The OAuth 2.0 Authorization Framework”

- “OpenID Connect Core 1.0” and most of its “standard claims”[12]
- RFC 7519 “JSON Web Token (JWT)” and RFC 7515 “JSON Web Signature (JWS)”
- RFC 8414 “OAuth 2.0 Authorization Server Metadata”

The server adopts the OpenID Connect discovery mechanism[13] using the “well-known” URI[14] constructed by concatenating the string “`/.well-known/openid-configuration`” to the Issuer’s URI. Therefore, in the initial version of the platform the full URI is https://aai.procancer-i.eu/.well-known/openid-configuration. The following is the returned JSON representation of the Identity Provider’s metadata, listing all the supported functionality and available endpoints:

```
{
    "response_types_supported": [
      "code"
    ],
    "introspection_endpoint":"https://aai.procancer-i.eu/oauth2/introspect",
    "grant_types_supported": [
      "authorization_code",
      "client_credentials",
      "refresh_token"
    ],
    "issuer": "https://aai.procancer-i.eu",
    "introspection_endpoint_auth_methods_supported": "none",
    "response_modes_supported": [
      "query"
    ],
    "claims_supported": [
      "iss", "sub", "aud",
      "iat","exp","jti", "name",
      "first_name", "family_name", "email"
    ],
    "subject_types_supported": [
      "public"
    ],
    "id_token_signing_alg_values_supported": [
```

[12] https://openid.net/specs/openid-connect-core-1_0.html#StandardClaims

[13] https://openid.net/specs/openid-connect-discovery-1_0.html

[14] Well-known URIs were introduced by RFC 5785 (https://datatracker.ietf.org/doc/html/rfc5785)

```
      "RS256"
    ],
    "code_challenge_methods_supported": [
      "S256"
    ],
    "token_endpoint_auth_methods_supported": [
      "client_secret_basic",
      "client_secret_post"
    ],
    "authorization_endpoint": "https://aai.procancer-i.eu/oauth2/auth",
    "userinfo_endpoint": "https://aai.procancer-i.eu/oauth2/userinfo",
    "end_session_endpoint": "https://aai.procancer-i.eu/logout",
    "token_endpoint": "https://aai.procancer-i.eu/oauth2/token",
    "jwks_uri": "https://aai.procancer-i.eu/oauth2/certs"
  }
```

Using the information above the client (an OpenID Connect “Relying Party”) can retrieve information about the various authorization endpoints as well as the public key of the issuer in order to verify the authenticity of the tokens issued.

5.2.2 User Authorization

The JSON Web Tokens issued by the Identity Provider carry group membership information for the authenticated user. The “wlcg.groups” claim conveys group membership (“roles”) about an authenticated end-user[15]. The claim value is an ordered JSON array of strings that contains the names of groups of which the user is a member in the context of the “virtual organization” (i.e., ProstateNet) that issued the Token. The currently available roles are:

- admins: users that are entitled to administer the system, assign new users to roles, etc.
- modellers: AI/ML model developers
- data-providers: users that augment the data platform with new datasets or other types of relevant data (e.g., the output of image processing tools)
- users: the generic role assigned to all authenticated users of the platform

As defined in Section 3.3 of the OAuth 2.0 specification, the client is permitted to specify the scope of the access request using the "scope" request parameter. In turn, the authorization

[15] M. Altunay et al., WLCG Common JWT Profiles. 2019. doi: 10.5281/zenodo.3460258.

server uses the "scope" response parameter to inform the client of the scope of the access token issued. The scope attribute is embedded in the access tokens issued by our Identity Provider and can be used to further control access to the platform's functionality. The currently supported scopes are qualified based on the component contacted (e.g. Image Store) and whether the client aims to retrieve ("read") or insert/update ("write") information. The following are some of the supported scopes:

- "dicomweb:read" : Search and retrieve image series and instances from the DICOM Image Store
- "dicomweb:write" : Upload new DICOM series (e.g. new patient data sets or image segmentations) to the DICOM Image Store
- "metadata:read" : Search and retrieve clinical data and their metadata from the Metadata Repository
- "metadata:write" : Upload new data and/or their metadata annotations into the

RBAC cannot cover complex relationships and constraints that are more granular than the role memberships. For example, an access control policy that restricts the management of the uploaded data only to the user that originally uploaded them cannot be handled with the coarse role-based access rules. In such situations, an Attribute Based Access Control (ABAC) approach, such as the one supported by XACML[16], is more relevant and will be pursued in the future versions of the platform. In preparation of this, we have adopted an ABAC approach for the implementation of the current, role-based, access control using the Open Policy Agent (OPA[17]) as the Policy Decision Point (PDP). OPA is a "cloud native" general purpose policy engine that can be used both as a library as well as a service for making authorization decisions. The policies (rules) in OPA are expressed in a declarative language inspired by Datalog called *Rego*. Below we have a test policy for allowing or denying access to a ProCAncer-I service based on the roles and the capabilities expressed through "scopes" defined above:

```
package prostatenet.rbac

# By default deny any access:
default allow = false
# A mapping between the roles (e.g. "modellers") and the list
# of permitted "scopes":
permissions = {
   # "Data providers" have both read and write access
   # to both data repositories:
   "data-providers": ["metadata:write", "metadata:read",
```

[16] http://www.oasis-open.org/committees/xacml/

[17] https://www.openpolicyagent.org/

```
                      "dicomweb:read", "dicomweb:write"],
   # "Modellers" have only read access:
   "modellers": ["metadata:read", "image-store:read"]
}

## In order to validate the signature of the JWTs
## read the Identity Provider's public key from an
## environment variable:
pubKey = opa.runtime().env.IDP_PUB_KEY

claims = payload {
   # Get the token from the input submitted by the service:
   bearer_token := input.access_token
   # Verify the signature on the Bearer token.
   io.jwt.verify_rs256(bearer_token, pubKey)

   # The `io.jwt.decode` function decodes the token and returns an
   # array:
   #
   #   [header, payload, signature]
   #
   # We extract the JWT payload using pattern matching:
   [_, payload, _] := io.jwt.decode(bearer_token)
}
grantedScopes = split(claims.scope, " ")

allow {
   # Only accept tokens for the ProstateNet audience (VO):
   claims.aud == "ProstateNet"

   # The scopes granted (contained in the JWT) should contain
   # the scope that the Service supplies through `input`:
   input.scope == grantedScopes[_]

   # Read the user's roles from the claims contained
   # in the JWT:
   groups := claims["wlcg.groups"]

   # The service's scope should be among the ones permitted
   # by any of the roles (groups) the user belongs to:
   permissions[groups[_]][_] == input.scope
}
```

The policy expressed above accepts as input a JWT access token encoding claims and the user's role membership information as well as the capability requested (for example "`metadata:write`" for updating the content of the Metadata Repository through its API). Based on the above policy description OPA will perform the following in order to find the value of the "allow" rule:

1) Validates the token based on the public key of the ProCAncer-I's Identity Provider. The public key, encoded in Privacy Enhanced Mail (PEM) format, is read from the IDP_PUB_KEY environment variable.
2) Decodes the token to get the "payload" that contains the various claims.
3) It retrieves the (potentially multiple) values of the granted "scopes" contained in the payload of the token and puts them in an array
4) It checks that the audience (the aud claim) is ProstateNet.
5) It checks that the capability requested by the service (e.g. the Metadata API) are indeed granted and contained in the scopes of the token.
6) It retrieves the user's role membership information contained in the wclg.groups claim.
7) It checks whether the requested capability through the input's scope attribute is contained in the permissions granted to one of the groups (roles) that the user belongs to.
8) If any check fails, the allow rule is "unified" to false, which is its default value.

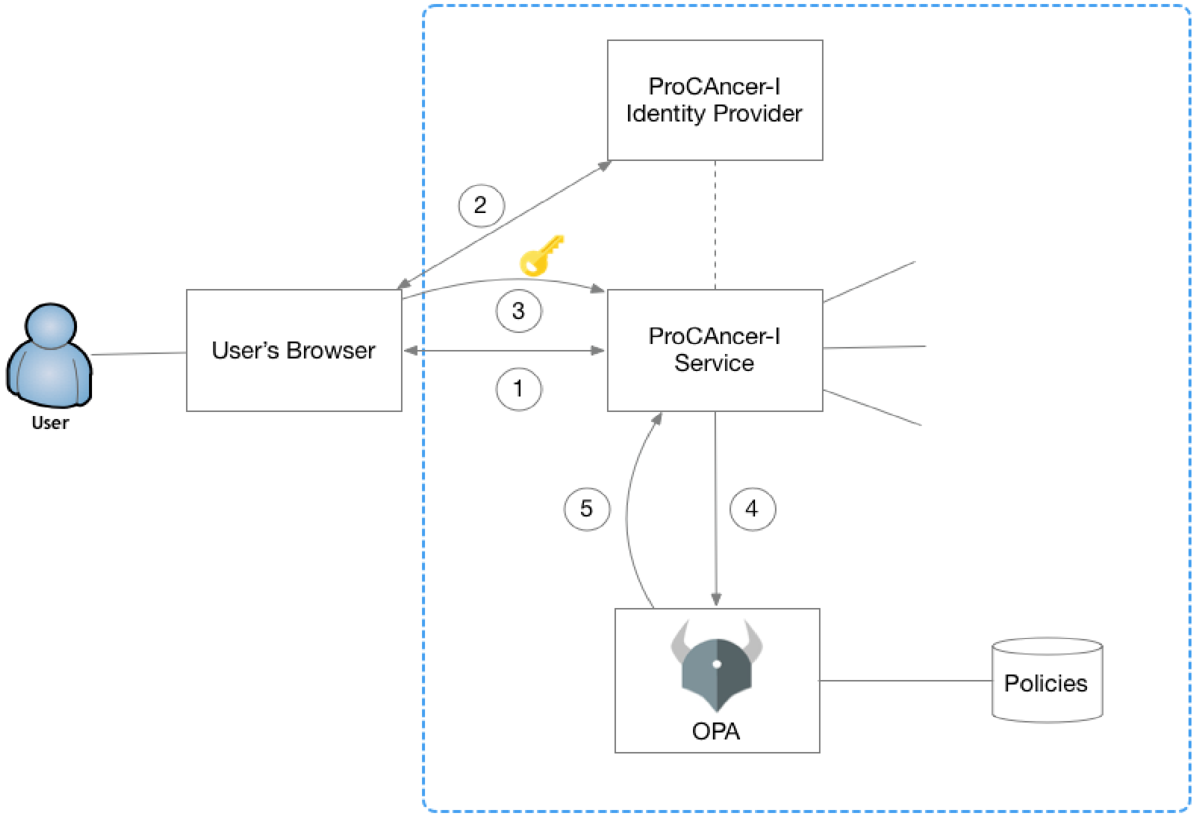


Figure 18. Making authorization decision using OPA

The whole process of authentication and authorization is shown in the relevant figure. The user visits a ProCAncer-I service (for example, the Metadata Repository) through their browser ①. If the user has not yet initiated a session with the service, s/he will be redirected to the Identity

Provider in order to authenticate ②. After the user authenticates (and implicitly authorizes the service to gain access to his/her profile and roles) s/he will be redirected back to the service, which will then be able to retrieve an access token ③. In order to make an authorization decision, such as whether the user is able to invoke a specific service functionality, the service contacts OPA supplying the token and the identification ("scope") of the requested functionality ④. OPA will then make a policy decision based on the established policies ⑤ and based on this decision the service will allow or deny the user's request ⑥. OPA therefore acts as the Policy Decision Point (PDP) while the service is a Policy Enforcement Point (PEP). Adopting OPA for making authorization decisions simplifies the architecture and leads to better security since the policies can be managed at a single place, while the decisions can be enforced in a distributed way, in the whole ProCAncer-I platform.

5.2.3 An indicative scenario for the use of Tokens and Single-Sign On

The following sequence diagram shows a typical interaction with core tools and APIs of the system and how the single-sign on and the token-based authentication works in practice. The platform itself is hosted on the cloud and therefore users use their web browser to access it. The main platform components shown are the following:

- The so called "Gateway" is the end user (web based) application that provides an overview of the imaging studies available in the platform
- The Annotation Tool is the web application for performing segmentations and other annotations on the user selected imaging data.
- The "AuthN/AuthZ" server is our Identity Provider that is linked with the ELIXIR AAI as described above
- The DICOM Store and the Metadata Repository are two APIs for accessing and managing the imaging and clinical data respectively.

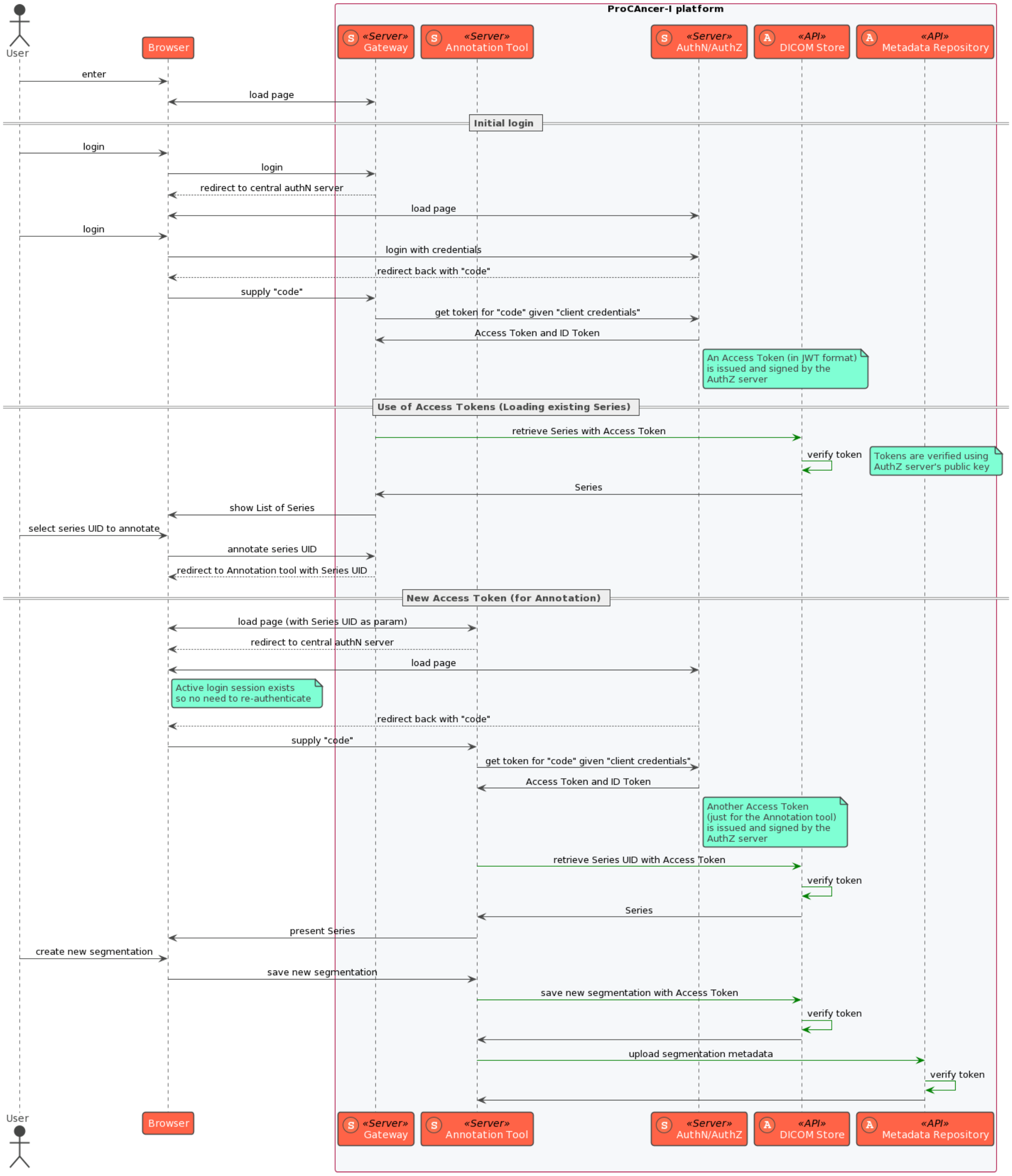

ProCAncer-I platform
User
Browser
«Server» Gateway
«Server» Annotation Tool
«Server» AuthN/AuthZ
«API» DICOM Store
«API» Metadata Repository
enter
load page
Initial login
login
login
redirect to central authN server
load page
login
login with credentials
redirect back with "code"
supply "code"
get token for "code" given "client credentials"
Access Token and ID Token
An Access Token (in JWT format) is issued and signed by the AuthZ server
Use of Access Tokens (Loading existing Series)
retrieve Series with Access Token
verify token
Tokens are verified using AuthZ server's public key
Series
show List of Series
select series UID to annotate
annotate series UID
redirect to Annotation tool with Series UID
New Access Token (for Annotation)
load page (with Series UID as param)
redirect to central authN server
load page
Active login session exists so no need to re-authenticate
redirect back with "code"
supply "code"
get token for "code" given "client credentials"
Access Token and ID Token
Another Access Token (just for the Annotation tool) is issued and signed by the AuthZ server
retrieve Series UID with Access Token
verify token
Series
present Series
create new segmentation
save new segmentation
save new segmentation with Access Token
verify token
upload segmentation metadata
verify token
User
Browser
«Server» Gateway
«Server» Annotation Tool
«Server» AuthN/AuthZ
«API» DICOM Store
«API» Metadata Repository

Figure 19. Interacting with the core ProCAncer-I components using token-based authentication

The user first visits the "Gateway" web page and, since s/he is unauthenticated, the browser is redirected in the "AuthN/AuthZ" server (Identity Provider) where using her/his ELIXIR credentials the user sign-ins as described above (not shown here). Next, according to the OAuth "authentication code" flow (defined in OAuth 2.0 RFC 6749, section 4.1[18]), the Identity Provider (IdP) redirects the browser back to the Gateway with a "code" parameter and finally the Gateway, using the supplied "code", retrieves an Access Token and, if requested, an ID Token that encodes user information[19]. Both tokens are in fact JSON Web Tokens (JWT) signed by the IdP's private key and have an expiration time[20].

After the retrieval of the access token, the Gateway server can make any API calls providing this as a "Bearer" token[21] such as the retrieval of imaging series from the DICOM Store. The DICOM Store offering the DICOMweb API verifies the token. At a minimal level, this verification checks that the token has indeed been signed by the IdP using IdP's public key and that it has not expired, but additional authorization checks can be performed using the user's identity and claims encoded in the access token with OPA as described above (not shown here). After the token's verification, the DICOM Store returns the imaging series to the Gateway and they are subsequently presented to the user.

When the user selects an imaging series to perform segmentation the Gateway redirects the browser to the Annotation Tool's web page. The annotation tool, in order to see whether the visitor is a real ProCAncer-I user, redirects user's browser again to the IdP web page, which is the single point for users' authentication in the platform. Since there's an active user session with the IdP, IdP generated immediately another code and redirects the browser back to the Annotation Tool without requesting the user to supply his/her credentials. The Annotation Tool exchanges the supplied code for another access token and an ID token. These JSON web tokens are totally new, with their own expiration time and encoding Annotation Tool's ID as the requesting entity. Using this new access token, the Annotation Tool makes API calls to the DICOM Store, in order to retrieve the input imaging series and upload the new segmentation file in the DICOM Seg format, and the Metadata Repository to upload the metadata associated with the segmentation task. These backend APIs verify the supplied access token as described above and if valid they perform the requested functionality.

The above scenario uses exclusively the "authorization code" grant of OAuth, which is its most advanced flow, since all the services, APIs, and tools mentioned above have a server-side component and are considered "confidential clients" by OAuth[22]. The IdP keeps a list of known clients and for each client a "secret" (password) and redirection URI is registered, so that no third

[18] https://datatracker.ietf.org/doc/html/rfc6749#section-4.1

[19] ID Tokens were introduced by OpenID Connect and "is an artifact that proves that the user has been authenticated". On the other hand, the access token allows the client (the "Gateway" in our example) to access specific resources (e.g., APIs) on behalf of the user.

[20] RFC9068 "JSON Web Token (JWT) Profile for OAuth 2.0 Access Tokens", https://datatracker.ietf.org/doc/html/rfc9068

[21] RFC6750 "The OAuth 2.0 Authorization Framework: Bearer Token Usage", https://datatracker.ietf.org/doc/html/rfc6750

[22] OAuth defines two client types, confidential and public, based on their ability to authenticate with the authorization server (i.e.,ability to maintain the confidentiality of their client credentials).

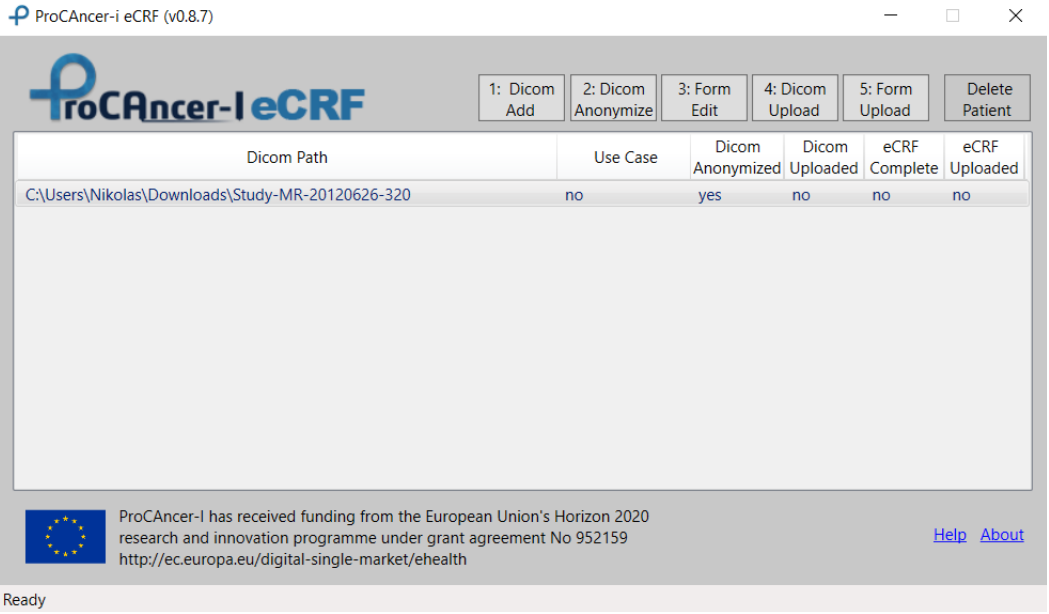
party (outside the platform) client can request user's authentication or the creation of access tokens. In the case of "public clients", such as Single Page Applications (SPAs) in Javascript, more provisions should be in place, but in the current version of the platform there are no such requirements.

5.3 Data Upload & eCRF

The ProCAncer-I eCRF – Data Upload Tool aims to support ProCAncer-I clinical partners in the process of compiling the required information and follow the defined protocols for uploading data to the project cloud repository. The tool integrates with the CTP anonymisation tool and with the ProCAncer-I repository services such as authentication, DICOM upload API Gateway and Metadata Catalog API, enabling the anonymisation and upload of data using methods that comply with privacy and security requirements. The tool is designed to follow a simple 5 steps workflow from the selection of the folder containing the DICOM files, to the anonymisation, edition of the clinical information and upload of DICOM and clinical information data. The ProCAncer-I eCRF – Data Upload Tool is a windows application that should be installed on Windows 10 or above computers with internet connection. After installation a shortcut icon can be available at the Start Menu and/or in the windows desktop.

5.3.1 eCRF – Data Upload Tool main window

The tool is designed to follow a simple workflow in 5 steps to compile and update the data to the ProCAncer-I cloud repositories:



Dicom Path	Use Case	Dicom Anonymized	Dicom Uploaded	eCRF Complete	eCRF Uploaded
C:\Users\Nikolas\Downloads\Study-MR-20120626-320	no	yes	no	no	no

ProCAncer-I has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952159 http://ec.europa.eu/digital-single-market/ehealth

Figure 20. eCRF – Data Upload Tool main window

1. Add the DICOM files;

2. Anonymise the DICOM files;
3. Edit the clinical information by selecting the appropriate Use Case;
4. Upload the DICOM files;
5. Upload the clinical information.

The workflow starts with "1: DICOM Add" for the selection of a case folder which contains the DICOM studies for the patient. The case folder should have been previously prepared with the studies relevant to the use case. If the use case requires multiple studies (baseline and follow-up) the user should create a folder for that patient and place there a copy of the DICOM folders of each relevant study. The PatientID of those DICOM studies must be the same and should not be empty.

Step 1: DICOM Add

After clicking on the button "1: DICOM Add", a new window will open (Figure 21) for selecting the appropriate folder. The selected folder or its subfolders must contain DICOM files, otherwise it will not be accepted.

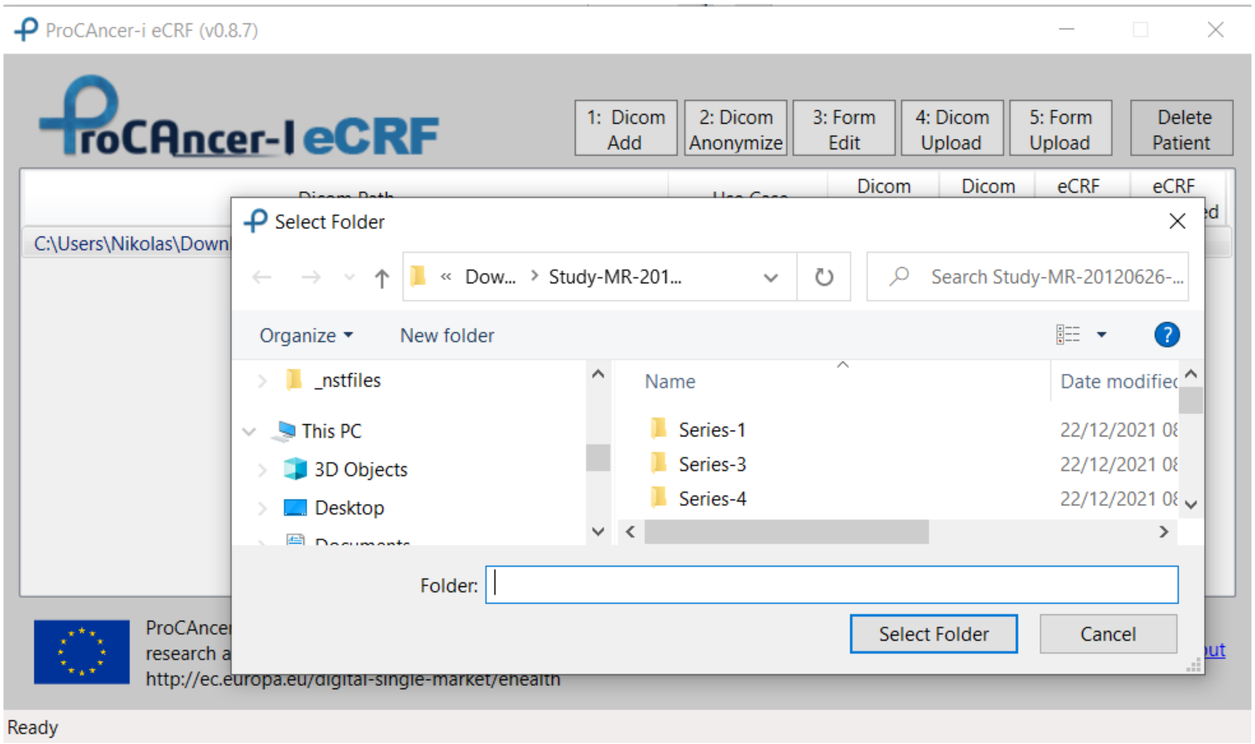
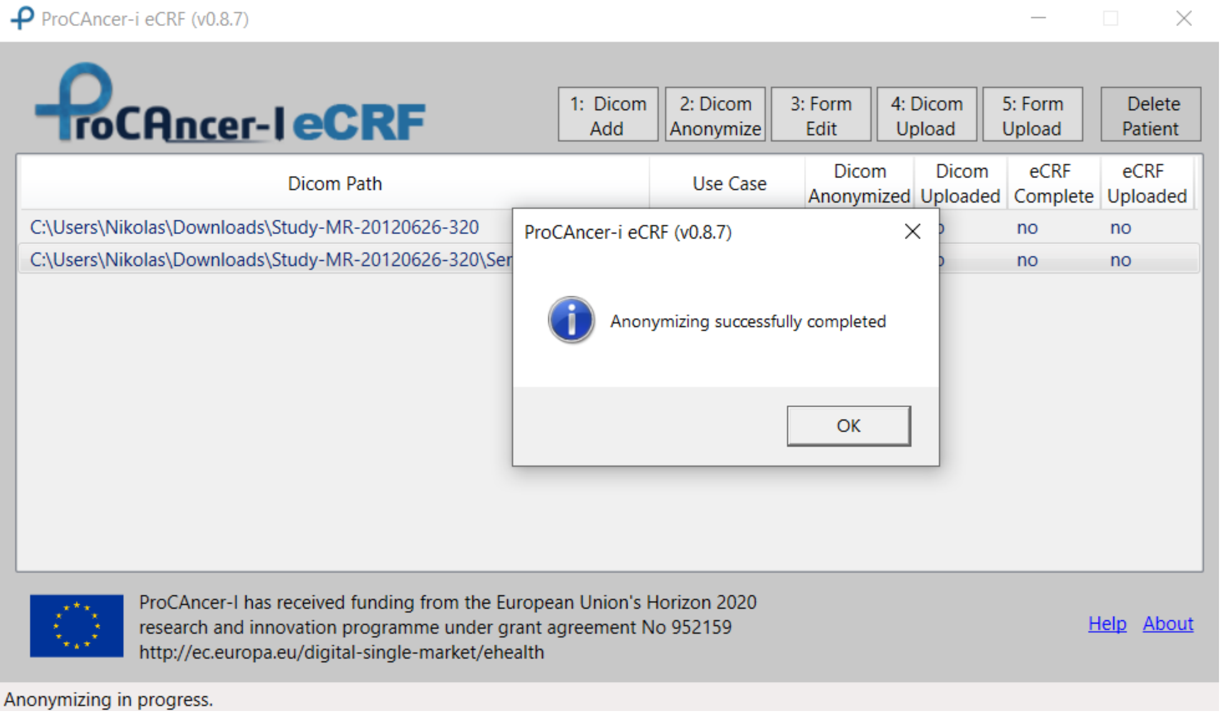


Figure 21. Selection of the DICOM folder for a new case

Once a new case folder is added, it will be shown in the list of cases below. Besides the DICOM folder the list shows the Use Case and the status of Anonymisation, DICOM Upload, eCRF completeness and eCRF Upload. These fields will be updated with the different steps in the workflow. To perform any of the following steps the user must select the case in the list.

Step 2: DICOM Anonymise

The second step “2: DICOM anonymise” runs the CTP anonymisation tool with the pre-defined anonymisation script that completely anonymises the DICOM files as defined in the project. By pressing the DICOM anonymise button the anonymisation process will start, showing the status of the process in the status area in the bottom of the page. At the end of the process a message will display the result of the anonymisation process, shown in Figure 22.



Dicom Path	Use Case	Dicom Anonymized	Dicom Uploaded	eCRF Complete	eCRF Uploaded
C:\Users\Nikolas\Downloads\Study-MR-20120626-320			o	no	no
C:\Users\Nikolas\Downloads\Study-MR-20120626-320\Ser			o	no	no

ProCAncer-i eCRF (v0.8.7)

Anonymizing successfully completed

ProCAncer-I has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952159 http://ec.europa.eu/digital-single-market/ehealth

Anonymizing in progress.

Figure 22. Anonymization step concluded

If the DICOM files were already previously anonymised, the process will show a message stating the DICOM folder was already anonymised.

ProCAncer – I Anonymisation strategy

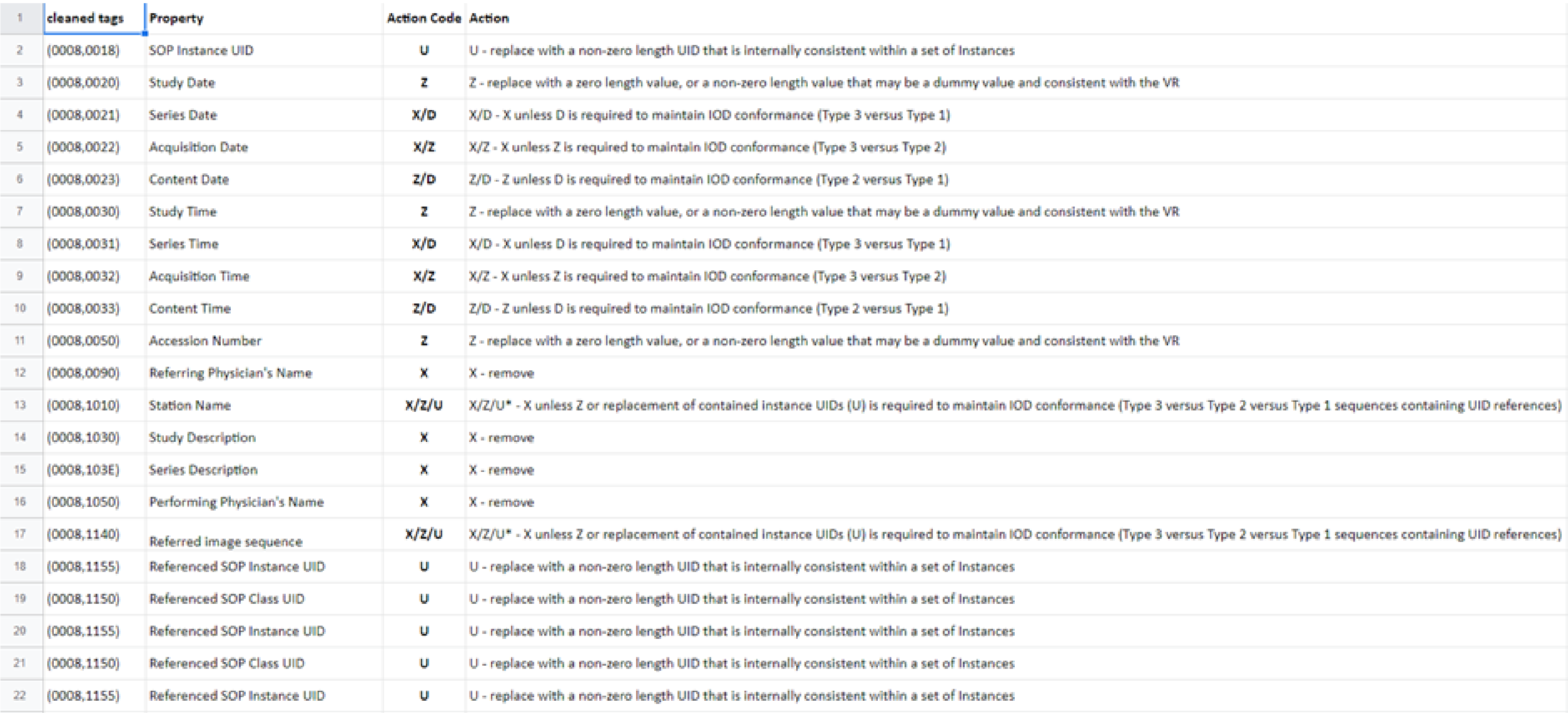
Anonymization is an irreversible processing operation which consists of using a set of techniques in such a way as to make it impossible, in practice, to identify the person by any means. This process is challenging when dealing with DICOM formatted data. The complexity lies in effectively performing an anonymization strategy while also preserving the value of the DICOM dataset as input for model development where the individual characteristics of each case have to be adequately described. Thus, the optimal compromise between the data quality and safety was sought extensively within the clinical and technical partners of the ProCAncer-I consortium.

During this process, all participants expressed their opinion based on their experiences as well as the active legislative framework of their country. During the discussions, a common working document was shared among all partners, which listed suggested DICOM tags containing Protected Health Information (PHI) from the DICOM Supplement on anonymization strategies. The two main pillars of the expected outcome were simplicity on one hand as not all ProCancer-I clinical partners were familiar with the anonymization process and efficiency on the other hand as the security and privacy issue is highly rated in ProCAncer-I priorities.

During the process for deciding the ProCAncer-I anonymization strategy, all members and the coordinating team examined a number of freely available solutions that the clinical partners could potentially adapt. The assessment included anonymization simulations with the identified tools, either with their default configurations or with the optimal set of parameters where this was an option. The results were assessed by the participating members of this anonymization task force and a second exercise was performed by the clinical partners who did not have internally an established anonymisation process or tool. All the tests were performed utilising phantom data that could be shared among the consortium without exposing real patient PHI.

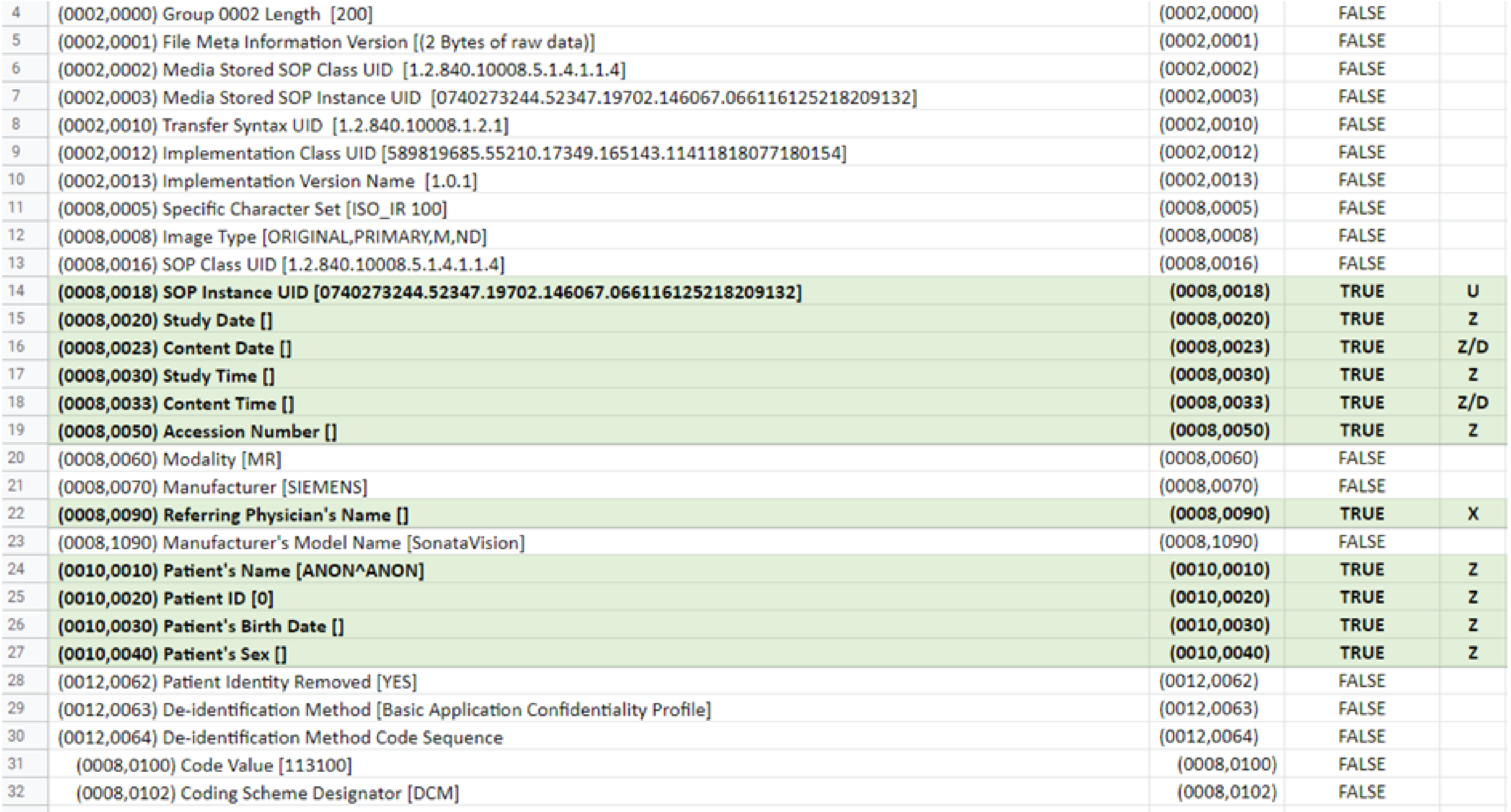
The total number of tags requiring modification were 38, and consequently tha evaluation of the selected tools were based on the number of PHI containing tags accurately modified by the default settings of each tool, as shown in the following figure.

Concerning other DICOM tags that were not included in DICOM Committee guidelines, an open discussion was held among partners where the participants were encouraged to share their opinions concerning the most appropriate modification of each tag. The main challenge is the large variability of MRI scanners and the software versions existing in the clinical sites, each of them producing a different version of the anonymized DICOM header list. Of utmost importance is the private DICOM tag list, which contains specific attributes in a non well-defined fashion from each vendor, creating high heterogeneity, even for datasets handled by the same tool and configuration. This heterogeneity was examined by gathering indicative results from each clinical partner that was willing to share it among the consortium.



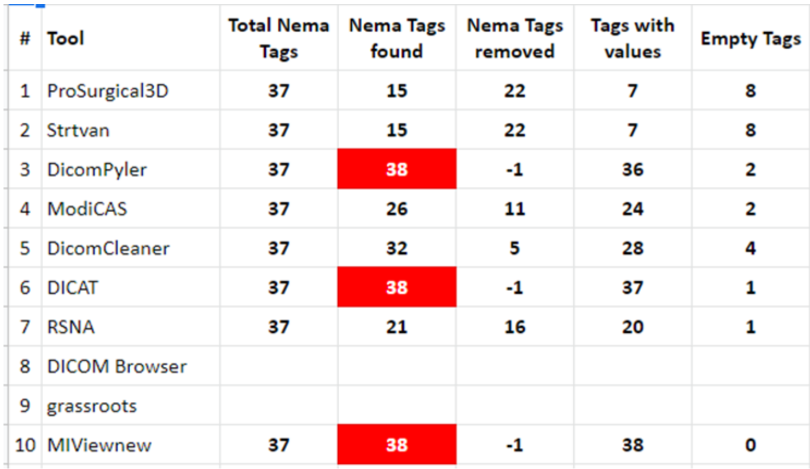
1	cleaned tags	Property	Action Code	Action
2	(0008,0018)	SOP Instance UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances
3	(0008,0020)	Study Date	Z	Z - replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
4	(0008,0021)	Series Date	X/D	X/D - X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
5	(0008,0022)	Acquisition Date	X/Z	X/Z - X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
6	(0008,0023)	Content Date	Z/D	Z/D - Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
7	(0008,0030)	Study Time	Z	Z - replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
8	(0008,0031)	Series Time	X/D	X/D - X unless D is required to maintain IOD conformance (Type 3 versus Type 1)
9	(0008,0032)	Acquisition Time	X/Z	X/Z - X unless Z is required to maintain IOD conformance (Type 3 versus Type 2)
10	(0008,0033)	Content Time	Z/D	Z/D - Z unless D is required to maintain IOD conformance (Type 2 versus Type 1)
11	(0008,0050)	Accession Number	Z	Z - replace with a zero length value, or a non-zero length value that may be a dummy value and consistent with the VR
12	(0008,0090)	Referring Physician's Name	X	X - remove
13	(0008,1010)	Station Name	X/Z/U	X/Z/U* - X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)
14	(0008,1030)	Study Description	X	X - remove
15	(0008,103E)	Series Description	X	X - remove
16	(0008,1050)	Performing Physician's Name	X	X - remove
17	(0008,1140)	Referred image sequence	X/Z/U	X/Z/U* - X unless Z or replacement of contained instance UIDs (U) is required to maintain IOD conformance (Type 3 versus Type 2 versus Type 1 sequences containing UID references)
18	(0008,1155)	Referenced SOP Instance UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances
19	(0008,1150)	Referenced SOP Class UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances
20	(0008,1155)	Referenced SOP Instance UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances
21	(0008,1150)	Referenced SOP Class UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances
22	(0008,1155)	Referenced SOP Instance UID	U	U - replace with a non-zero length UID that is internally consistent within a set of Instances

Figure 23. Indicative proposed actions for selected PHI containing tags



4	(0002,0000) Group 0002 Length [200]	(0002,0000)	FALSE	
5	(0002,0001) File Meta Information Version [(2 Bytes of raw data)]	(0002,0001)	FALSE	
6	(0002,0002) Media Stored SOP Class UID [1.2.840.10008.5.1.4.1.1.4]	(0002,0002)	FALSE	
7	(0002,0003) Media Stored SOP Instance UID [0740273244.52347.19702.146067.066116125218209132]	(0002,0003)	FALSE	
8	(0002,0010) Transfer Syntax UID [1.2.840.10008.1.2.1]	(0002,0010)	FALSE	
9	(0002,0012) Implementation Class UID [589819685.55210.17349.165143.11411818077180154]	(0002,0012)	FALSE	
10	(0002,0013) Implementation Version Name [1.0.1]	(0002,0013)	FALSE	
11	(0008,0005) Specific Character Set [ISO_IR 100]	(0008,0005)	FALSE	
12	(0008,0008) Image Type [ORIGINAL,PRIMARY,M,ND]	(0008,0008)	FALSE	
13	(0008,0016) SOP Class UID [1.2.840.10008.5.1.4.1.1.4]	(0008,0016)	FALSE	
14	**(0008,0018) SOP Instance UID [0740273244.52347.19702.146067.066116125218209132]**	**(0008,0018)**	**TRUE**	**U**
15	**(0008,0020) Study Date []**	**(0008,0020)**	**TRUE**	**Z**
16	**(0008,0023) Content Date []**	**(0008,0023)**	**TRUE**	**Z/D**
17	**(0008,0030) Study Time []**	**(0008,0030)**	**TRUE**	**Z**
18	**(0008,0033) Content Time []**	**(0008,0033)**	**TRUE**	**Z/D**
19	**(0008,0050) Accession Number []**	**(0008,0050)**	**TRUE**	**Z**
20	(0008,0060) Modality [MR]	(0008,0060)	FALSE	
21	(0008,0070) Manufacturer [SIEMENS]	(0008,0070)	FALSE	
22	**(0008,0090) Referring Physician's Name []**	**(0008,0090)**	**TRUE**	**X**
23	(0008,1090) Manufacturer's Model Name [SonataVision]	(0008,1090)	FALSE	
24	**(0010,0010) Patient's Name [ANON^ANON]**	**(0010,0010)**	**TRUE**	**Z**
25	**(0010,0020) Patient ID [0]**	**(0010,0020)**	**TRUE**	**Z**
26	**(0010,0030) Patient's Birth Date []**	**(0010,0030)**	**TRUE**	**Z**
27	**(0010,0040) Patient's Sex []**	**(0010,0040)**	**TRUE**	**Z**
28	(0012,0062) Patient Identity Removed [YES]	(0012,0062)	FALSE	
29	(0012,0063) De-identification Method [Basic Application Confidentiality Profile]	(0012,0063)	FALSE	
30	(0012,0064) De-identification Method Code Sequence	(0012,0064)	FALSE	
31	(0008,0100) Code Value [113100]	(0008,0100)	FALSE	
32	(0008,0102) Coding Scheme Designator [DCM]	(0008,0102)	FALSE	

Figure 24. Indicative testing process, showing whether the selected software tool successfully (True/False) performed the action proposed by the DICOM Anonymization Committee. This evaluation was repeated for a large number of freely available software tools.



#	Tool	Total Nema Tags	Nema Tags found	Nema Tags removed	Tags with values	Empty Tags
1	ProSurgical3D	37	15	22	7	8
2	Strtvan	37	15	22	7	8
3	DicomPyler	37	38	-1	36	2
4	ModiCAS	37	26	11	24	2
5	DicomCleaner	37	32	5	28	4
6	DICAT	37	38	-1	37	1
7	RSNA	37	21	16	20	1
8	DICOM Browser					
9	grassroots					
10	MIViewnew	37	38	-1	38	0

Figure 25. Software tool evaluation method, based on the number of successfully modified PHI containing DICOM tags

The final decision for the ProCAncer-I anonymization strategy was to follow ***a two-stage anonymization process***. At the first stage, all partners are allowed to use the software tool of their preference with the restriction that it must be capable to successfully remove/modify the number of PHI tags as defined by the DICOM guidelines. At this stage, the issue of DICOM list heterogeneity still exists but the datasets are successfully stripped out by any PHI containing tags. The second stage of anonymization addresses the problem of DICOM tag list heterogeneity (different MRI scanner vendors and software versions), and the existence of the private tag list. This process is performed by the well-known RSNA utility embedded in the ProCAncer-I, after specific configuration performed by FORTH and B3D to search and retrieve useful private DICOM tags and then horizontally select the least number of useful and common among all partners DICOM tags that will be kept for the next stages of the project.

Summarizing in ProCAncer-I the consortium agreed to adopt both the “blacklisting” and the ‘whitelisting’ anonymisation strategy. As a first step only PHI containing data were modified or removed inside each clinical partner’s premises by the software tool of their own. At the second stage, during the upload process an automated procedure running the second whitelisting process is performed, keeping the least number of necessary and commonly existing DICOM tags ensuring homogeneity among the DICOM tag list among all partners.

Step 3: Form Edit

The third step “3: Form Edit” allows the user to select for each patient a particular form according to the available clinical, imaging, pathology, treatment, follow-up information. Since a single patient could be useful in more than one single Use Case, we optimized data collection in the eCRF, to avoid duplicates and time-consuming data entry by clinical partners. Adopting this solution, each patient is assigned a Form and a specific label, defining the Use Cases in which that particular patient could be used. The edition of the clinical information can be made several times until its completion. The Form Edit button will display a new window containing the different Forms, as detailed in the following sub-chapter. The user's manual contains a specific chapter to describe the required information for each Form. After saving the information on the selected Form, this window will close, and the case will be updated on the list and labeled with the appropriate Use Cases. Once the clinical information is complete, the case can be uploaded to the ProCAncer-I cloud repository, to the staging area.

Step 4: DICOM upload

The fourth step "4: DICOM upload" allows to upload the anonymised DICOM files of the selected case. It will only upload files if the case was already anonymised and have the clinical information complete. This operation requires internet connection to ProCAncer-I APIs. By pressing the DICOM upload button the status bar will show the progress of the upload process. At the end of the process the status bar will show the result of the upload process.

Step 5: Form Upload

The final step in the workflow is "5: Form Upload" that allows to upload the clinical information of the selected case to the ProCAncer-I cloud repository. It will only upload the information if the case was already anonymised and the clinical information is complete. This operation requires internet connection to ProCAncer-I APIs. By pressing the eCRF upload button the status bar will show the progress of the upload process. At the end of the process the status bar will show the result of the upload process.

DICOM Remove

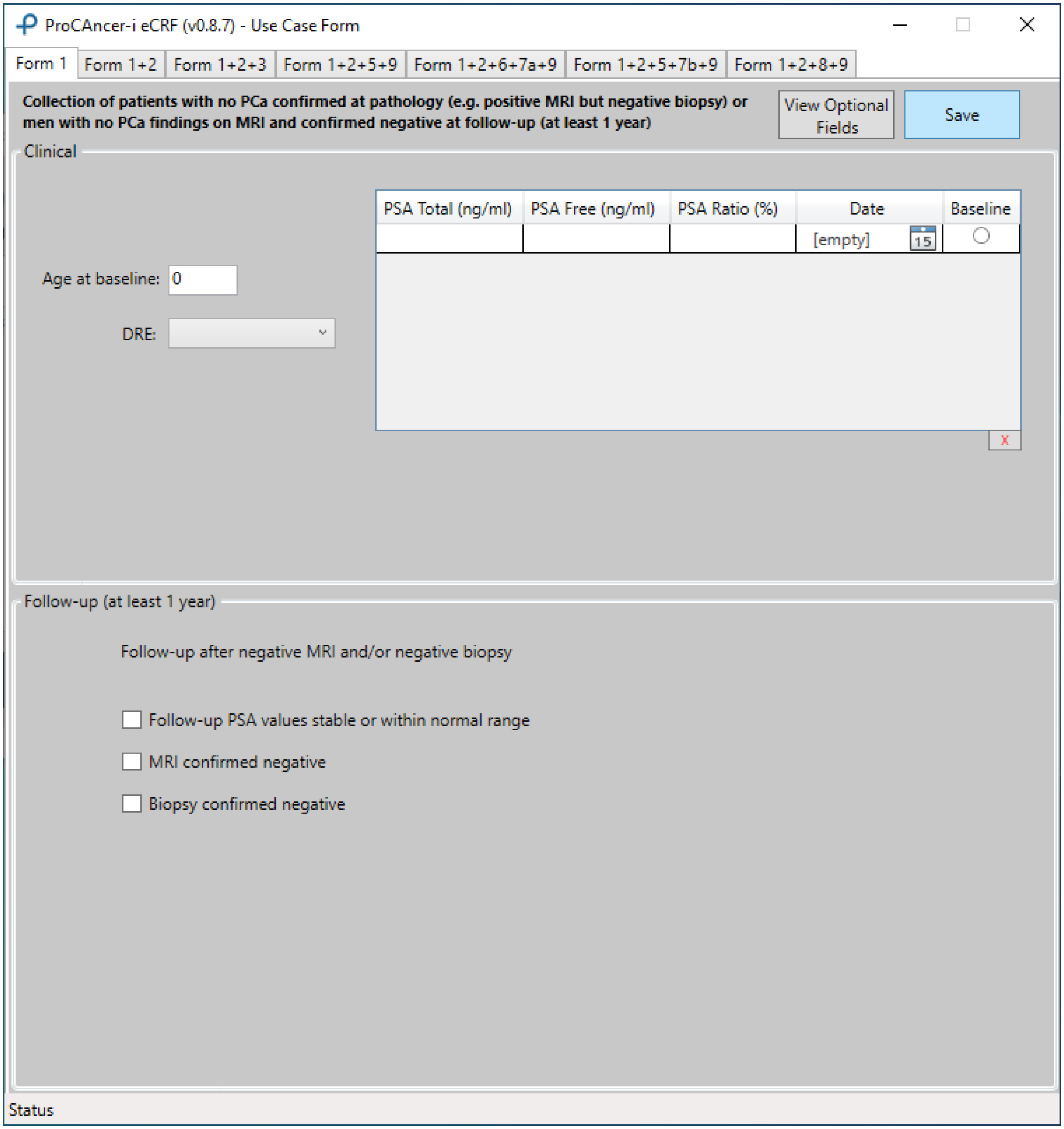
The user may remove a case from the list of cases by pressing the DICOM Remove button. This function will clear all the information created on the eCRF tool on the local machine related to that case. The original DICOM folder will remain. If the case was already uploaded the information on the cloud related to the case will also remain.

5.3.2 eCRF Forms

The eCRF Forms enables the user to select different Forms according to the imaging and non-imaging data available for each patient. Each Form contains different sections to be filled with clinical, imaging, pathology, treatment and follow-up information. Each tab will contribute to form a specific subgroup of patients which will be used to create the final dataset for one or more Use Cases. By pressing the save button the form will save the information on the currently selected form and return to the main application window.

eCRF Form 1

Collection of patients with no PCa confirmed at pathology (e.g. positive MRI but negative biopsy) or men with no PCa findings on MRI and confirmed negative at follow-up (at least 1 year). This form will contain only subjects representing the negative (control) group for Use Case 1 (detection of PCa) and will contain both true negative MRI cases and false positive MRI examinations. This form has two information sections (Figure 26): Clinical and Follow-up.



Collection of patients with no PCa confirmed at pathology (e.g. positive MRI but negative biopsy) or men with no PCa findings on MRI and confirmed negative at follow-up (at least 1 year)

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty]	○

Follow-up (at least 1 year)

Follow-up after negative MRI and/or negative biopsy

☐ Follow-up PSA values stable or within normal range

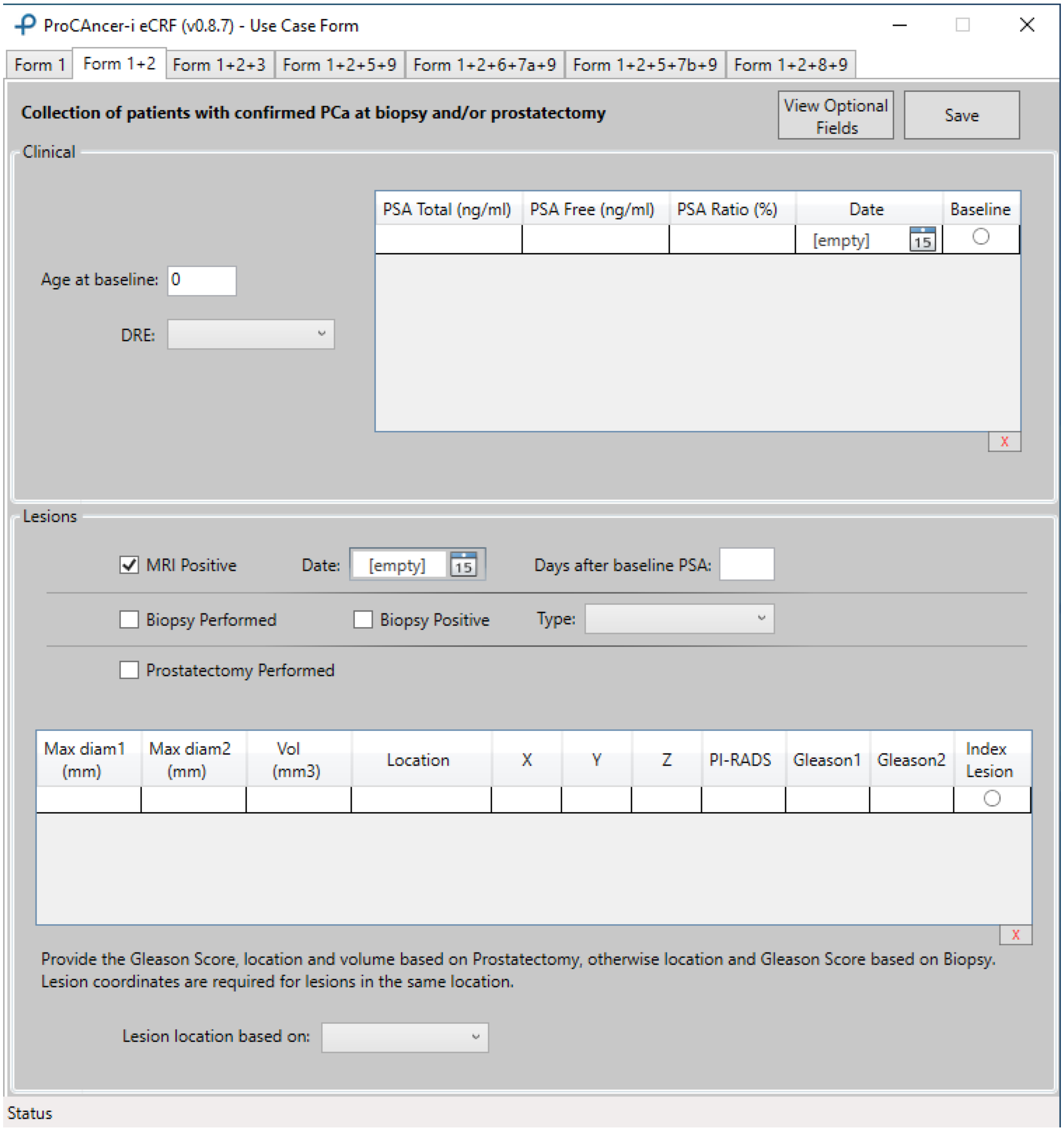
☐ MRI confirmed negative

☐ Biopsy confirmed negative

Figure 26. Form 1

eCRF Form 1 + 2

Collection of patients with confirmed PCa at biopsy and/or prostatectomy. This form will collect all patients who underwent pathological assessment, with confirmed presence of PCa and Gleason Score evaluation. These patients will represent the positive group in Use Case 1 (detection of PCa) and they will also be used in Use Case 2 (characterization of aggressive/non aggressive cancer) since all Gleason Score data will be collected. This form has two information sections (Figure 27): Clinical and Lesions.



Collection of patients with confirmed PCa at biopsy and/or prostatectomy

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty] 15	○

Lesions

☑ MRI Positive Date: [empty] 15 Days after baseline PSA:

□ Biopsy Performed □ Biopsy Positive Type:

□ Prostatectomy Performed

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Provide the Gleason Score, location and volume based on Prostatectomy, otherwise location and Gleason Score based on Biopsy.
Lesion coordinates are required for lesions in the same location.

Lesion location based on:

Figure 27. Form 1+2

Lesion Location map

The lesion location uses the 36 positions from PI-RADS:

- The right and left peripheral zones (PZ) at prostate base, midgland, and apex are each subdivided into three sections: anterior (a), medial posterior (mp), and lateral posterior (lp).
- The right and left transition zones (TZ) at prostate base, midgland, and apex are each subdivided into two sections: anterior (a) and posterior (p).
- The central zone (CZ) is included in the prostate base around the ejaculatory ducts.
- The anterior fibromuscular stroma (AS) is divided into right/left at the prostate base, midgland, and apex.

The user should check all the locations that apply for the selected lesion and click the save button to save and return to the previous window. There is a reset button to uncheck all locations and start again.

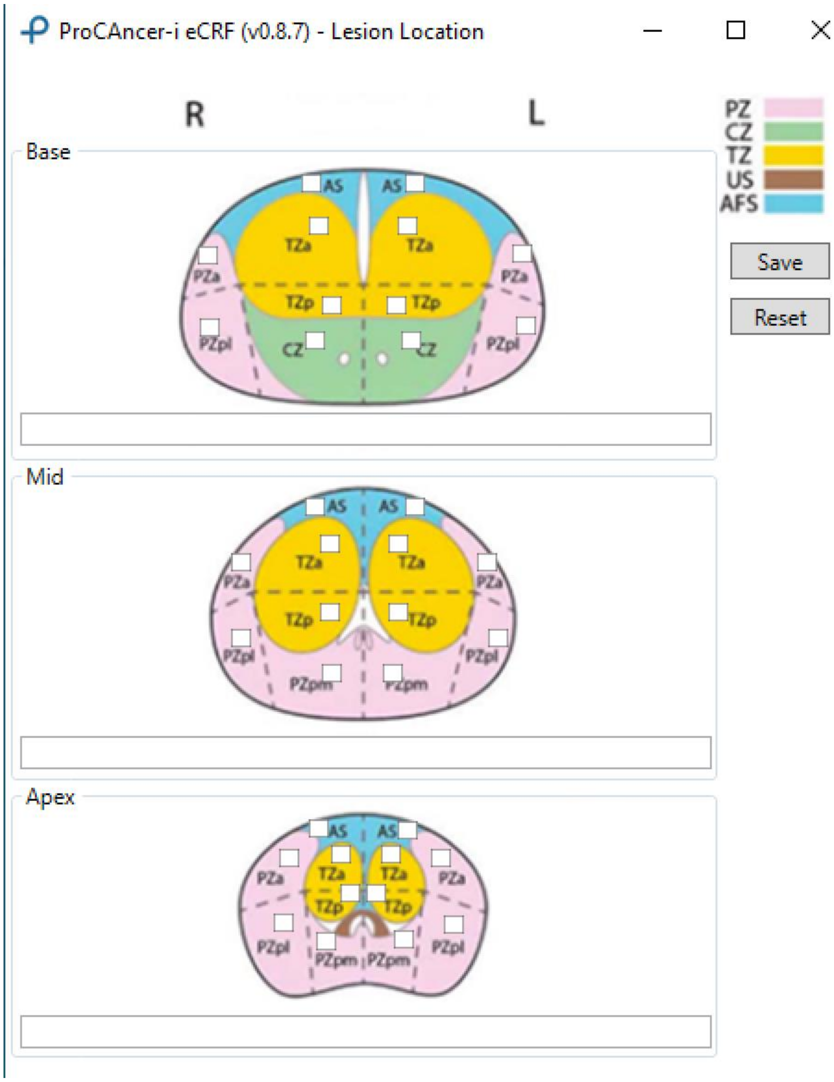
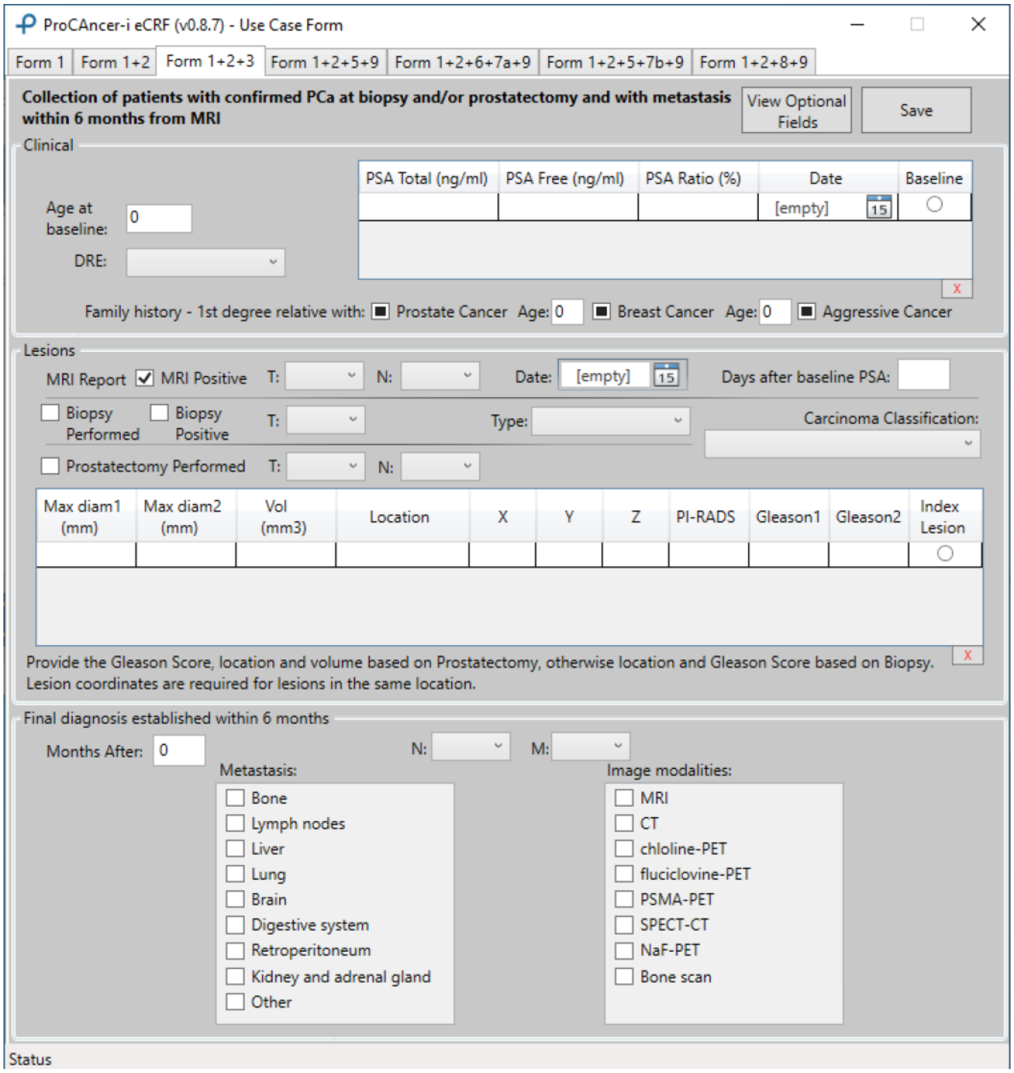


Figure 28. Lesion location map form

eCRF Form 1 + 2 + 3

Collection of patients with confirmed PCa at biopsy and/or prostatectomy and with metastasis within 6 months from MRI. This form will be used to collect patients with known metastatic cancer and will represent the positive group for Use Case 3 (metastatic PCa). All patients collected in this Form will also be used to populate Use Case 1 (detection of PCa) and Use Case 2 (characterization of aggressive/non-aggressive cancer). This form has three information sections (Figure 29): Clinical, Lesions and Final diagnosis.



Collection of patients with confirmed PCa at biopsy and/or prostatectomy and with metastasis within 6 months from MRI

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty] 15	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report ☑ MRI Positive T: N: Date: [empty] 15 Days after baseline PSA:

□ Biopsy Performed □ Biopsy Positive T: Type: Carcinoma Classification:

□ Prostatectomy Performed T: N:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Provide the Gleason Score, location and volume based on Prostatectomy, otherwise location and Gleason Score based on Biopsy.
Lesion coordinates are required for lesions in the same location.

Final diagnosis established within 6 months

Months After: 0 N: M:

Metastasis:
- □ Bone
- □ Lymph nodes
- □ Liver
- □ Lung
- □ Brain
- □ Digestive system
- □ Retroperitoneum
- □ Kidney and adrenal gland
- □ Other

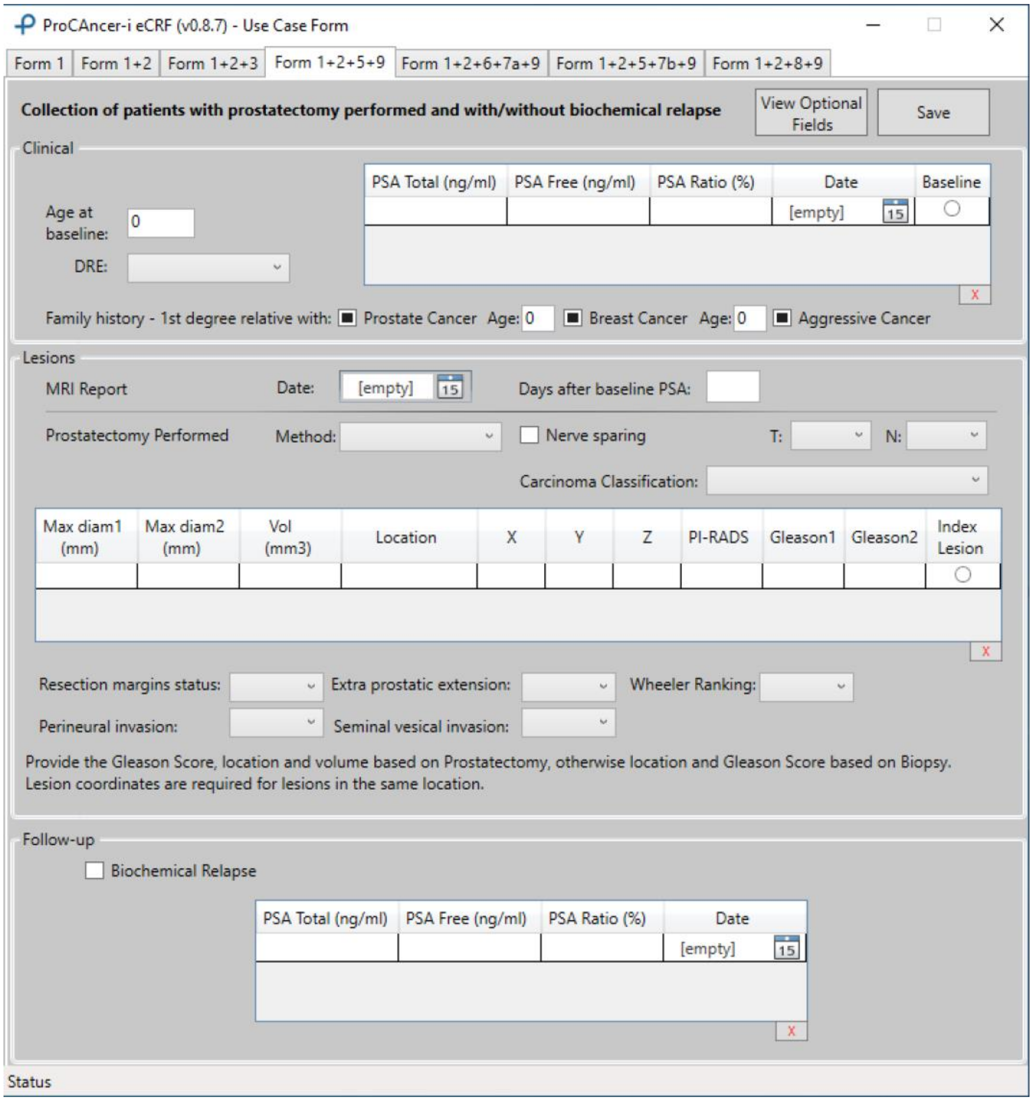
Image modalities:
- □ MRI
- □ CT
- □ chloline-PET
- □ fluciclovine-PET
- □ PSMA-PET
- □ SPECT-CT
- □ NaF-PET
- □ Bone scan

Status

Figure 29. Form 1+2+3

eCRF Form 1 + 2 + 5 + 9

Collection of patients with prostatectomy performed and with/without biochemical relapse. This Form will be useful for multiple Use Cases, as it will collect all patients who underwent radical prostatectomy. Therefore data will be used for: i) Use Case 1 (detection of PCa) since all patients in this Form will have positive pathological results; ii) Use Case 2 (characterization of aggressive/non aggressive cancer) since all patients will have a Gleason score assessment; iii) Use Case 5 (biochemical relapse/non relapse after prostatectomy) because in this form we will collect information about several parameters that could predict the likelihood of relapse after treatment; iv) Use Case 9 (detection of PCa and selection of best treatment), which is the last scenario that will be implemented to predict the best treatment strategy according to baseline MRI.This form has three information sections (Figure 30): Clinical, Lesions and Follow-up.



ProCAncer-i eCRF (v0.8.7) - Use Case Form

Collection of patients with prostatectomy performed and with/without biochemical relapse

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty] 15	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report Date: [empty] 15 Days after baseline PSA:

Prostatectomy Performed Method: ☐ Nerve sparing T: N:

Carcinoma Classification:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Resection margins status: Extra prostatic extension: Wheeler Ranking:

Perineural invasion: Seminal vesical invasion:

Provide the Gleason Score, location and volume based on Prostatectomy, otherwise location and Gleason Score based on Biopsy.
Lesion coordinates are required for lesions in the same location.

Follow-up

☐ Biochemical Relapse

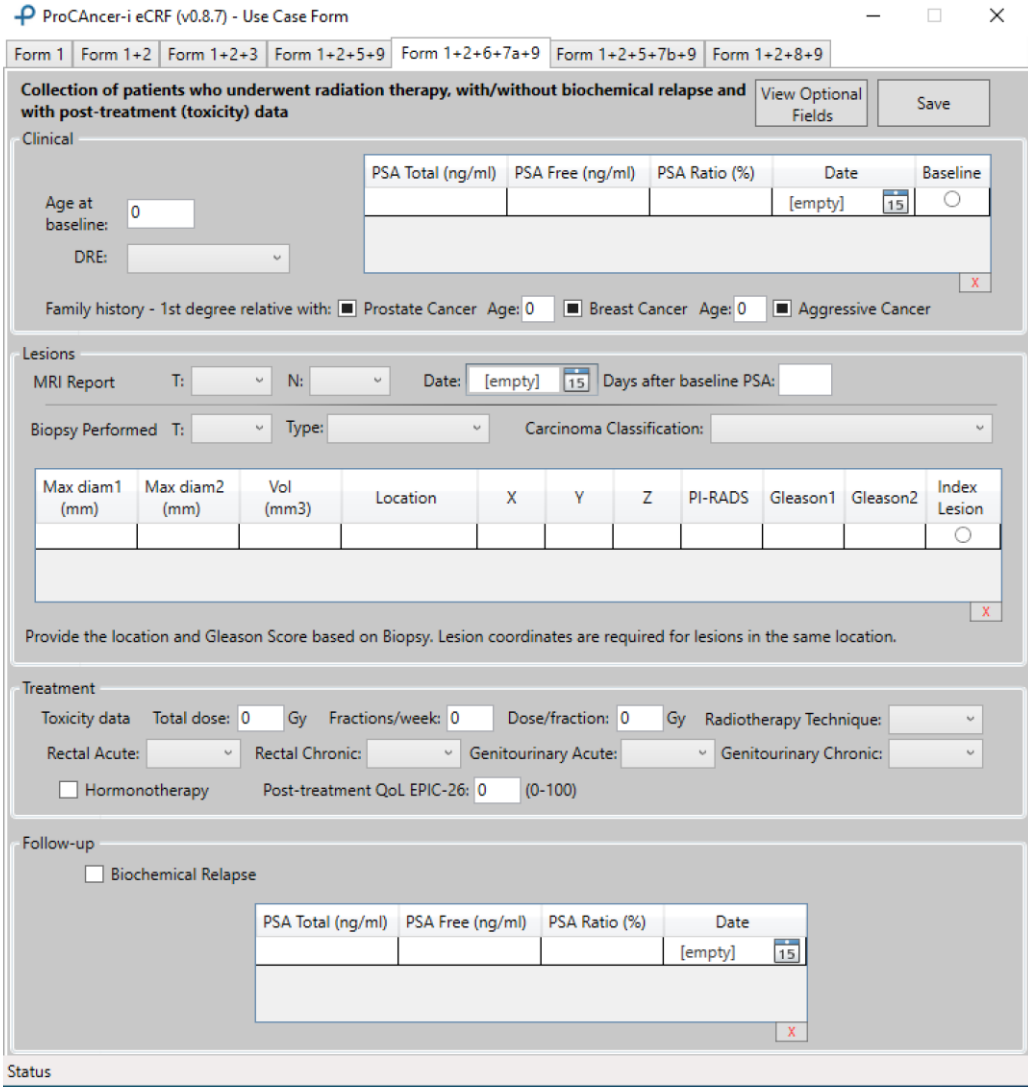
PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date
			[empty] 15

Status

Figure 30. Form 1+2+5+9

eCRF Form 1 + 2 + 6 + 7a + 9

Collection of patients who underwent radiation therapy, with/without biochemical relapse and with post-treatment (toxicity) data. This form is very similar to the previous one, but it is related to patients undergoing radiation therapy. It will be contribute to: Use Case 1 (detection of PCa) since all patients treated with radiation therapy and collected in this Form will have at least one positive PCa finding on pathology; ii) Use Case 2 (characterization of aggressive/non aggressive cancer) since all patients will have a Gleason score assessment; iii) Use Case 6 (biochemical relapse/non relapse after radiation therapy) because this form will collect information about PSA post-treatment; iv) Use Case 7a (side effects after radiation therapy) since the Form will collect parameters related to the radiation treatment and also post-treatment toxicity data; v) Use Case 9 (detection of PCa and selection of best treatment), which is the last scenario we will implement to predict the best treatment according to baseline MRI. This form has four information sections (Figure 31): Clinical, Lesions, Treatment and Follow-up.



Collection of patients who underwent radiation therapy, with/without biochemical relapse and with post-treatment (toxicity) data

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty]	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report T: N: Date: [empty] Days after baseline PSA:

Biopsy Performed T: Type: Carcinoma Classification:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Provide the location and Gleason Score based on Biopsy. Lesion coordinates are required for lesions in the same location.

Treatment

Toxicity data Total dose: 0 Gy Fractions/week: 0 Dose/fraction: 0 Gy Radiotherapy Technique:

Rectal Acute: Rectal Chronic: Genitourinary Acute: Genitourinary Chronic:

□ Hormonotherapy Post-treatment QoL EPIC-26: 0 (0-100)

Follow-up

□ Biochemical Relapse

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date
			[empty]

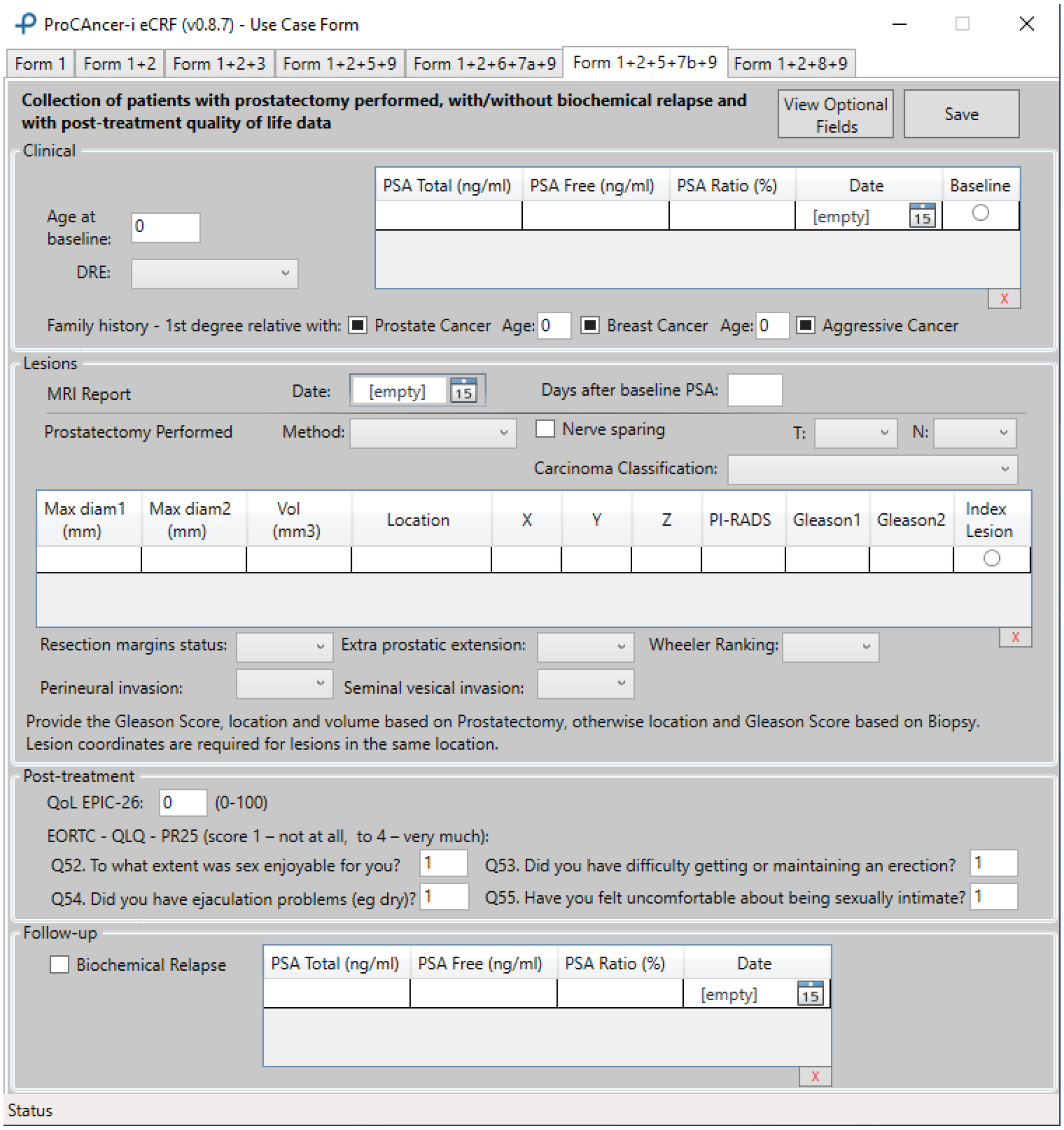
Status

Figure 31. Form 1+2+6+7a+9

eCRF Form 1 + 2 + 5 + 7b + 9

Collection of patients with prostatectomy performed, with/without biochemical relapse and with post-treatment quality of life data. This Form is very similar to Form 1+2+5+9 with the addition of post-treatment and quality of life assessment. It will useful for multiple Use Cases, which are:

i) Use Case 1 (detection of PCa) since all patients collected with this Form will have positive pathological results; ii) Use Case 2 (characterization of aggressive/non aggressive cancer) since all patients will have Gleason score assessment; iii) Use Case 5 (biochemical relapse/non relapse after prostatectomy) because in this form we will collect information about several parameters that could predict the likelihood of relapse; iv) Use Case 7b (side effects after radical prostatectomy) since the Form provide post-treatment data related to urinary incontinence, irritative/obstructive bowel, sexual, hormonal domains and erectile dysfunction, which could severely limit the quality of life of treated patients; v) Use Case 9 (detection of PCa and selection of best treatment), which is the last scenario we will implement to predict the best treatment according to baseline MRI. This form has four information sections (Figure 32): Clinical, Lesions, Post-treatment and Follow-up.



ProCAncer-i eCRF (v0.8.7) - Use Case Form

Form 1 | Form 1+2 | Form 1+2+3 | Form 1+2+5+9 | Form 1+2+6+7a+9 | Form 1+2+5+7b+9 | Form 1+2+8+9

Collection of patients with prostatectomy performed, with/without biochemical relapse and with post-treatment quality of life data

View Optional Fields | Save

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty] 15	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report Date: [empty] 15 Days after baseline PSA:

Prostatectomy Performed Method: □ Nerve sparing T: N:

Carcinoma Classification:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Resection margins status: Extra prostatic extension: Wheeler Ranking:

Perineural invasion: Seminal vesical invasion:

Provide the Gleason Score, location and volume based on Prostatectomy, otherwise location and Gleason Score based on Biopsy. Lesion coordinates are required for lesions in the same location.

Post-treatment

QoL EPIC-26: 0 (0-100)

EORTC - QLQ - PR25 (score 1 – not at all, to 4 – very much):

Q52. To what extent was sex enjoyable for you? 1 Q53. Did you have difficulty getting or maintaining an erection? 1

Q54. Did you have ejaculation problems (eg dry)? 1 Q55. Have you felt uncomfortable about being sexually intimate? 1

Follow-up

□ Biochemical Relapse

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date
			[empty] 15

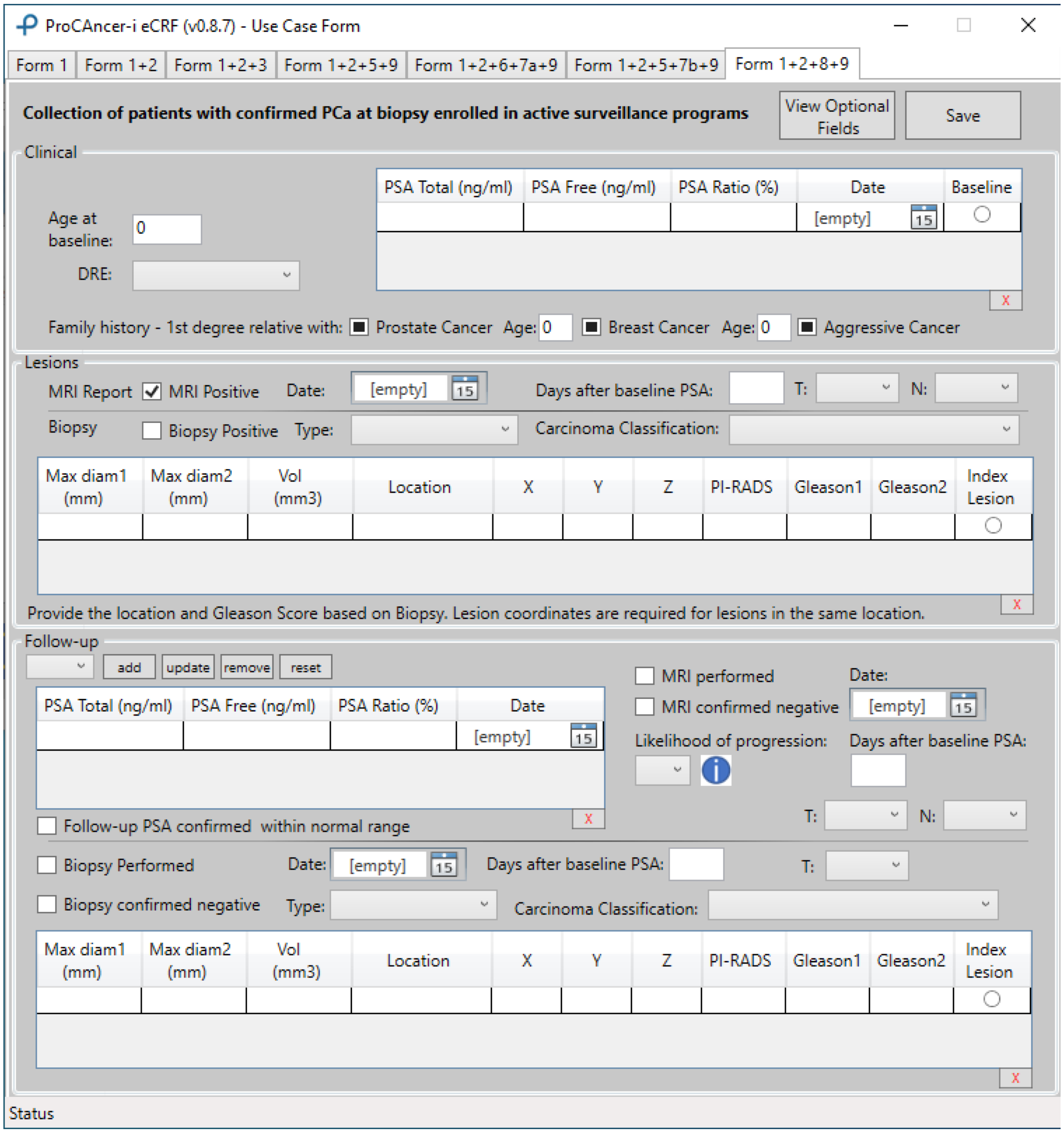
Status

Figure 32. Form 1+2+5+7b+9

eCRF Form 1 + 2 + 8 + 9

Collection of patients with confirmed PCa at biopsy enrolled in active surveillance programs .

This last form will collect data from biopsy and therefore it will be used to populate both Use Case 1 (detection of PCa) and Use Case 2 (characterization of aggressive/non aggressive cancer). In addition, the follow-up data will be used for Use Case 8, to predict the risk of disease progression in patients who are undergoing active surveillance. At last, all information from this Form will be used also for Use Case 9, (detection of PCa and selection of best treatment), in which the best treatment strategy will be predicted, according to baseline MRI. This form has three information sections (Figure 33): Clinical, Lesions and Follow-up (which can have multiple records).



Collection of patients with confirmed PCa at biopsy enrolled in active surveillance programs

Clinical

Age at baseline: 0

DRE:

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date	Baseline
			[empty]	○

Family history - 1st degree relative with: ■ Prostate Cancer Age: 0 ■ Breast Cancer Age: 0 ■ Aggressive Cancer

Lesions

MRI Report ☑ MRI Positive Date: [empty] Days after baseline PSA: T: N:

Biopsy □ Biopsy Positive Type: Carcinoma Classification:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Provide the location and Gleason Score based on Biopsy. Lesion coordinates are required for lesions in the same location.

Follow-up

PSA Total (ng/ml)	PSA Free (ng/ml)	PSA Ratio (%)	Date
			[empty]

□ MRI performed Date: [empty]

□ MRI confirmed negative

Likelihood of progression: Days after baseline PSA:

T: N:

□ Follow-up PSA confirmed within normal range

□ Biopsy Performed Date: [empty] Days after baseline PSA: T:

□ Biopsy confirmed negative Type: Carcinoma Classification:

Max diam1 (mm)	Max diam2 (mm)	Vol (mm3)	Location	X	Y	Z	PI-RADS	Gleason1	Gleason2	Index Lesion
										○

Figure 33. Form U1+2+8+9

5.3.3 ProCAncer-I upload process decision flowchart

As described in previous sections of this deliverable in ProCAncer-I the consortium has identified and described in detail various clinical UCs. The developed eCRF tools incorporates various tabs/forms where local data controllers will upload data (imaging and clinical). These data are linked with the various identified UCs. In supporting the transparency of the data uploading process and in order to reduce human errors, a decision flowchart has been prepared to assist end-users to understand which form(s) of the tool should be filled. The flowchart is shown in Figure 34.

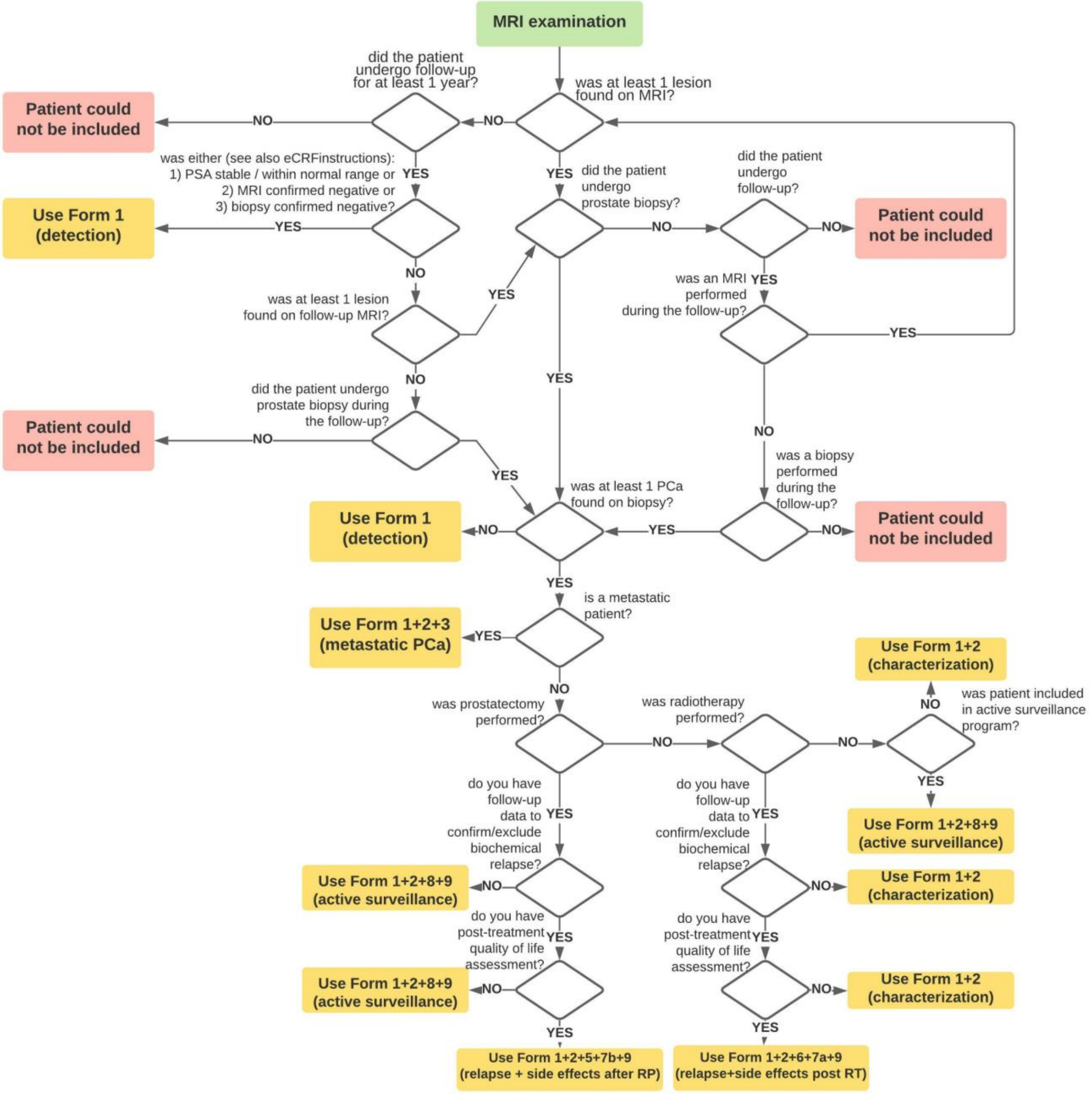
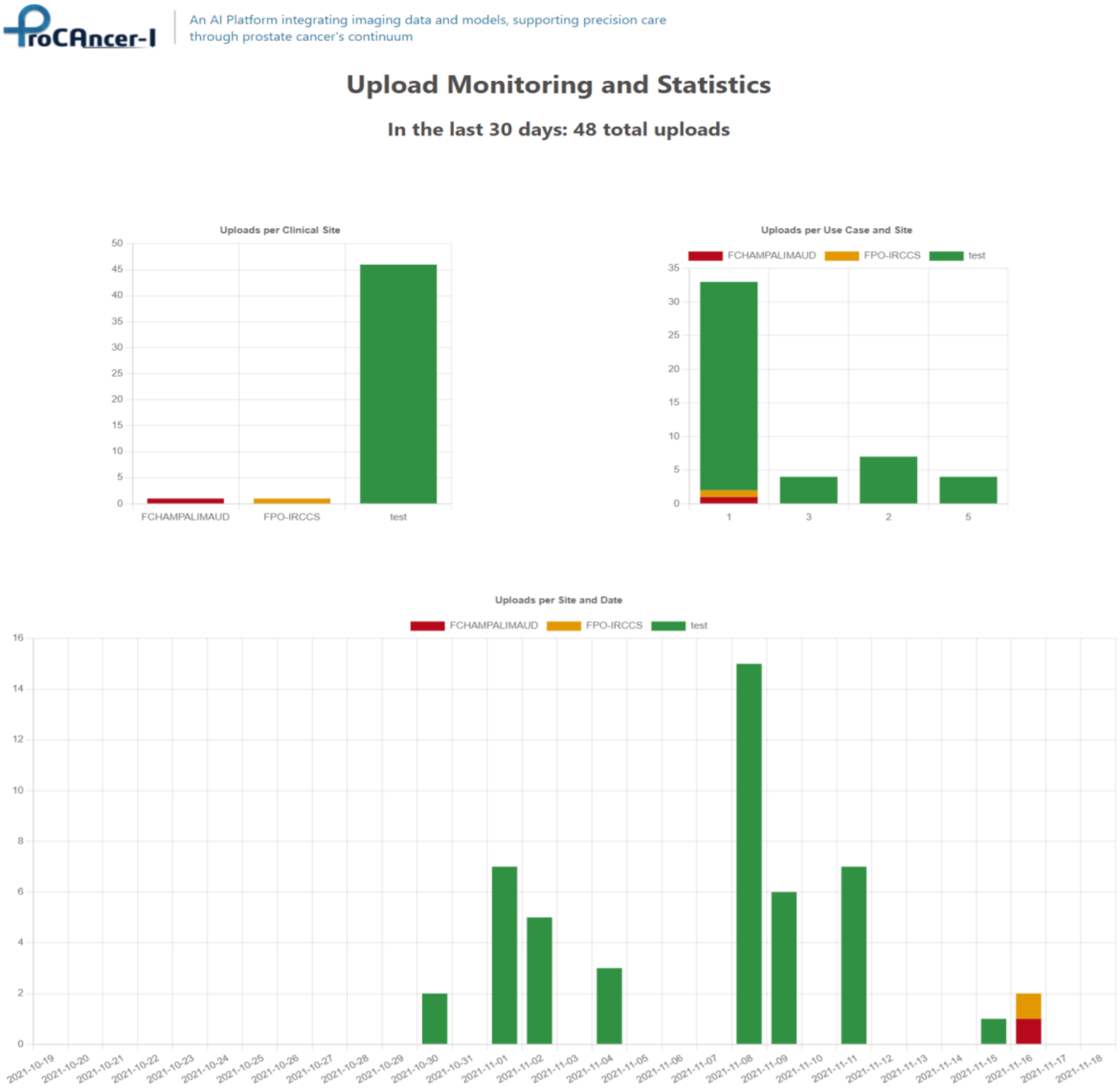

MRI examination
was at least 1 lesion found on MRI?
did the patient undergo follow-up for at least 1 year?
NO
NO
Patient could not be included
was either (see also eCRFinstructions): 1) PSA stable / within normal range or 2) MRI confirmed negative or 3) biopsy confirmed negative?
YES
YES
did the patient undergo prostate biopsy?
did the patient undergo follow-up?
Use Form 1 (detection)
YES
NO
NO
Patient could not be included
NO
YES
was at least 1 lesion found on follow-up MRI?
YES
was an MRI performed during the follow-up?
YES
NO
YES
did the patient undergo prostate biopsy during the follow-up?
Patient could not be included
NO
NO
YES
was a biopsy performed during the follow-up?
was at least 1 PCa found on biopsy?
Use Form 1 (detection)
NO
YES
NO
Patient could not be included
YES
is a metastatic patient?
Use Form 1+2+3 (metastatic PCa)
YES
Use Form 1+2 (characterization)
NO
NO
was patient included in active surveillance program?
was prostatectomy performed?
was radiotherapy performed?
NO
NO
YES
do you have follow-up data to confirm/exclude biochemical relapse?
YES
do you have follow-up data to confirm/exclude biochemical relapse?
YES
Use Form 1+2+8+9 (active surveillance)
Use Form 1+2+8+9 (active surveillance)
NO
NO
Use Form 1+2 (characterization)
do you have post-treatment quality of life assessment?
YES
do you have post-treatment quality of life assessment?
YES
Use Form 1+2+8+9 (active surveillance)
NO
NO
Use Form 1+2 (characterization)
YES
YES
Use Form 1+2+5+7b+9 (relapse + side effects after RP)
Use Form 1+2+6+7a+9 (relapse+side effects post RT)

Figure 34. Upload process decision flowchart

5.4 Monitoring tools for the data upload phase

In ProCAncer-I based on the DoA there will be two phases for uploading data in the provided cloud infrastructure. One for uploading retrospective MRI imaging case studies along with the necessary clinical data and metadata and one for prospective one. In order to monitor and assess the uploading the consortium has developed two main services. These are apart from the tools that the platform will integrate for exploring the rich dataset that will reside and be stored in the ProstateNet.

Specifically, the following screenshot presents a visual report for the case studies that are uploaded from the different clinical sites which are part of the project. Based on the provided reports the coordination team and the responsible partners for uploading the data could assess the process and monitor it in order to achieve the targeted numbers and prepare quickly mitigation plans if any foreseen issue arises



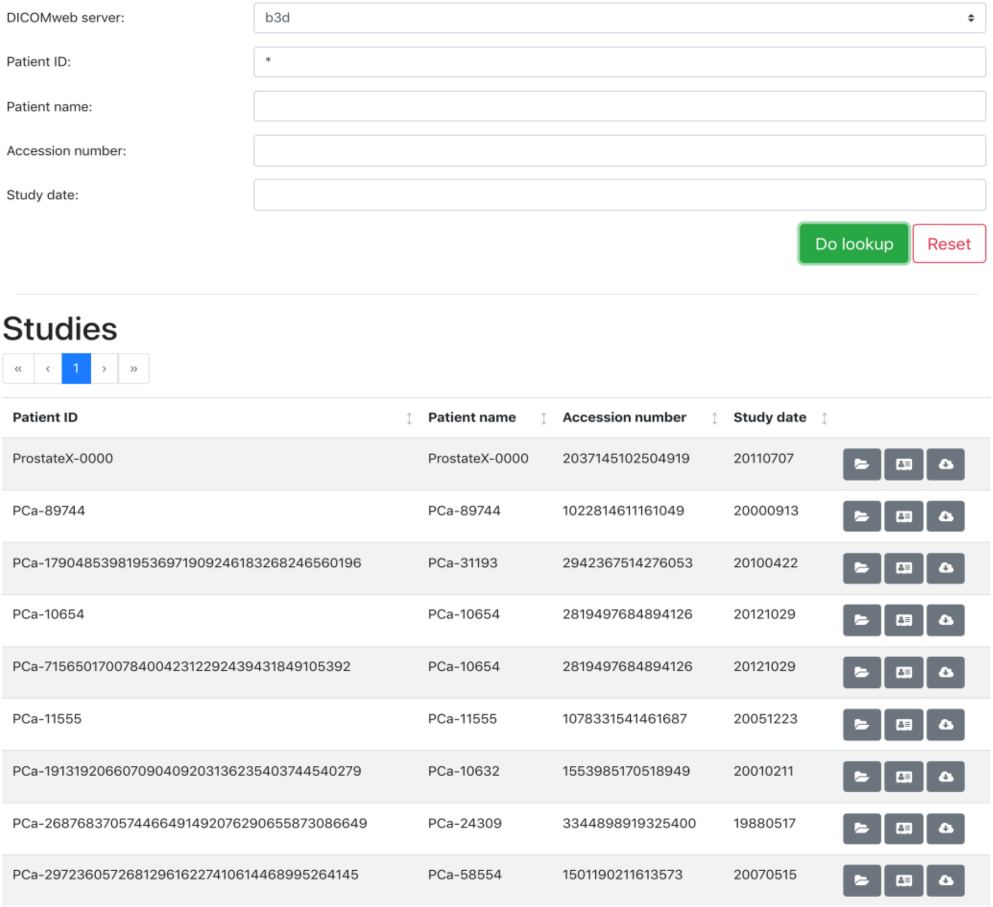
ProCAncer-I | An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum

Upload Monitoring and Statistics

In the last 30 days: 48 total uploads

Figure 35. Screenshot for the visual reporting of the data upload process.

In addition, utilising the provided DICOMserver API a second service has been integrated where the partners could assess and monitor the imaging uploaded data as depicted in the following figure.



DICOMweb server: b3d

Patient ID: *

Patient name:

Accession number:

Study date:

Studies

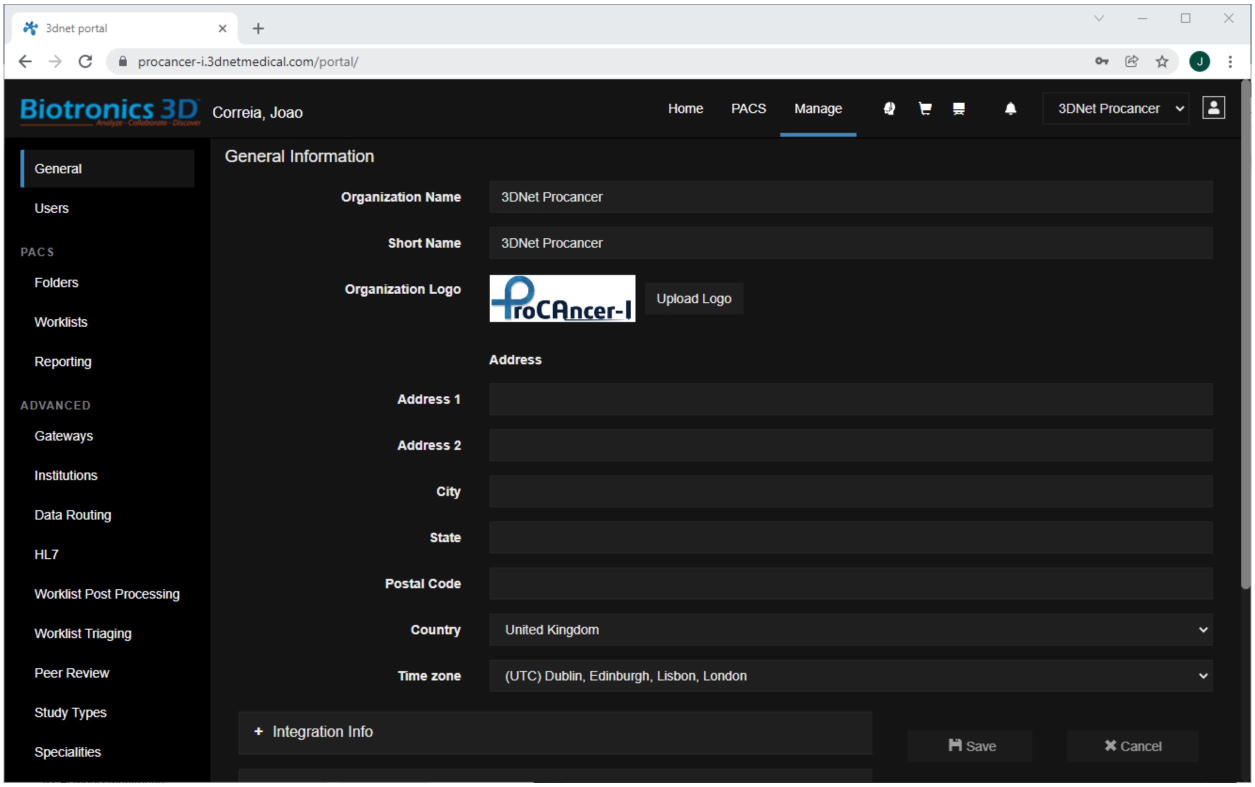
Patient ID	Patient name	Accession number	Study date	
ProstateX-0000	ProstateX-0000	2037145102504919	20110707	
PCa-89744	PCa-89744	1022814611161049	20000913	
PCa-179048539819536971909246183268246560196	PCa-31193	2942367514276053	20100422	
PCa-10654	PCa-10654	2819497684894126	20121029	
PCa-71565017007840042312292439431849105392	PCa-10654	2819497684894126	20121029	
PCa-11555	PCa-11555	1078331541461687	20051223	
PCa-191319206607090409203136235403744540279	PCa-10632	1553985170518949	20010211	
PCa-268768370574466491492076290655873086649	PCa-24309	3344898919325400	19880517	
PCa-297236057268129616227410614468995264145	PCa-58554	1501190211613573	20070515	

Figure 36. Screenshot for the visual reporting of the imaging data upload process.

5.5 Data Sharing

The image repository supports the sharing of the datasets through the implementation of roles, organisations and folders for imaging data and for annotations. Users will be allowed to share their datasets, images and annotations, with specific users or all users of the platform. Each organisation can contain different folders and worklists. Users have specific roles in organisations and access rights defined by folder and worklist. Specific users with the role of OrgAdmin for an organisation can create and manage folders, worklists and users.

The figures below illustrate the management of the organization (Figure 37), the management user access to folders (Figure 38) and the management of worklists (Figure 39).



Correia, Joao

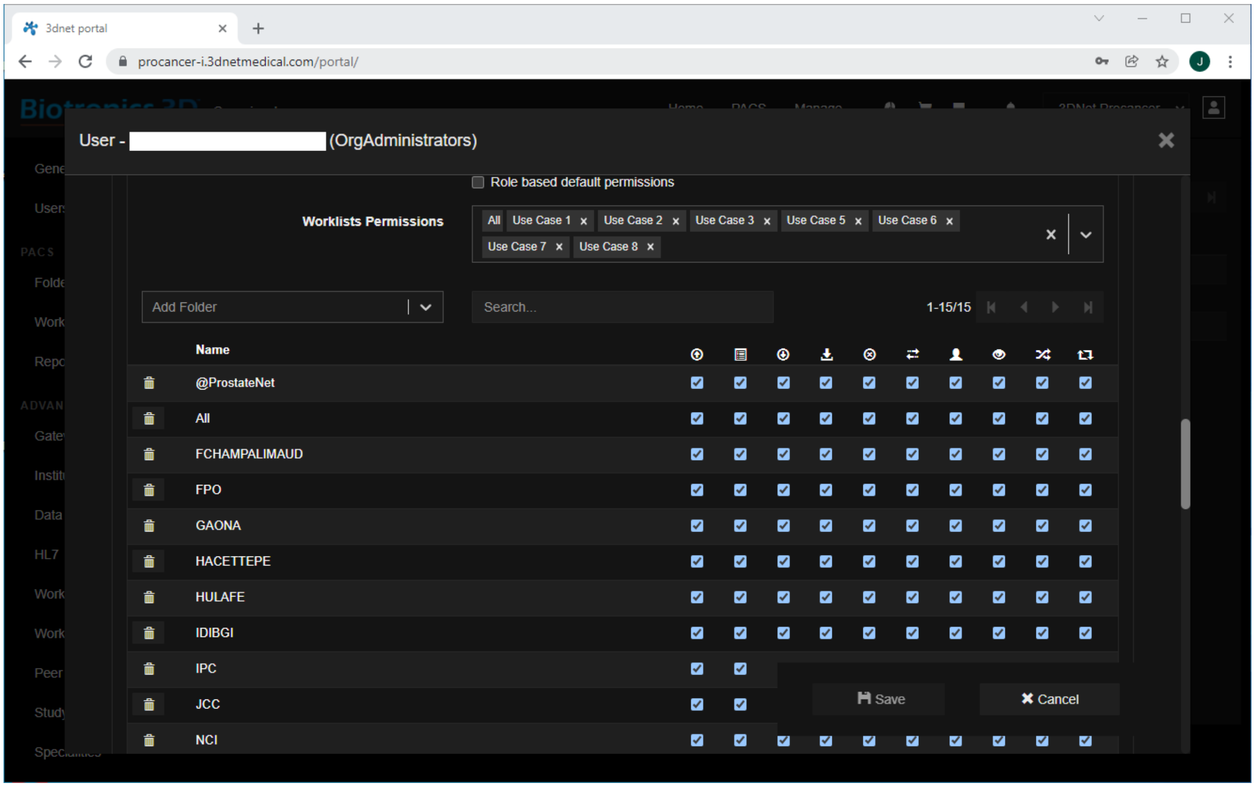
General Information

Organization Name	3DNet Procancer
Short Name	3DNet Procancer
Organization Logo	ProCAncer-I

Address

Address 1	
Address 2	
City	
State	
Postal Code	
Country	United Kingdom
Time zone	(UTC) Dublin, Edinburgh, Lisbon, London

Figure 37. Management of an organisation



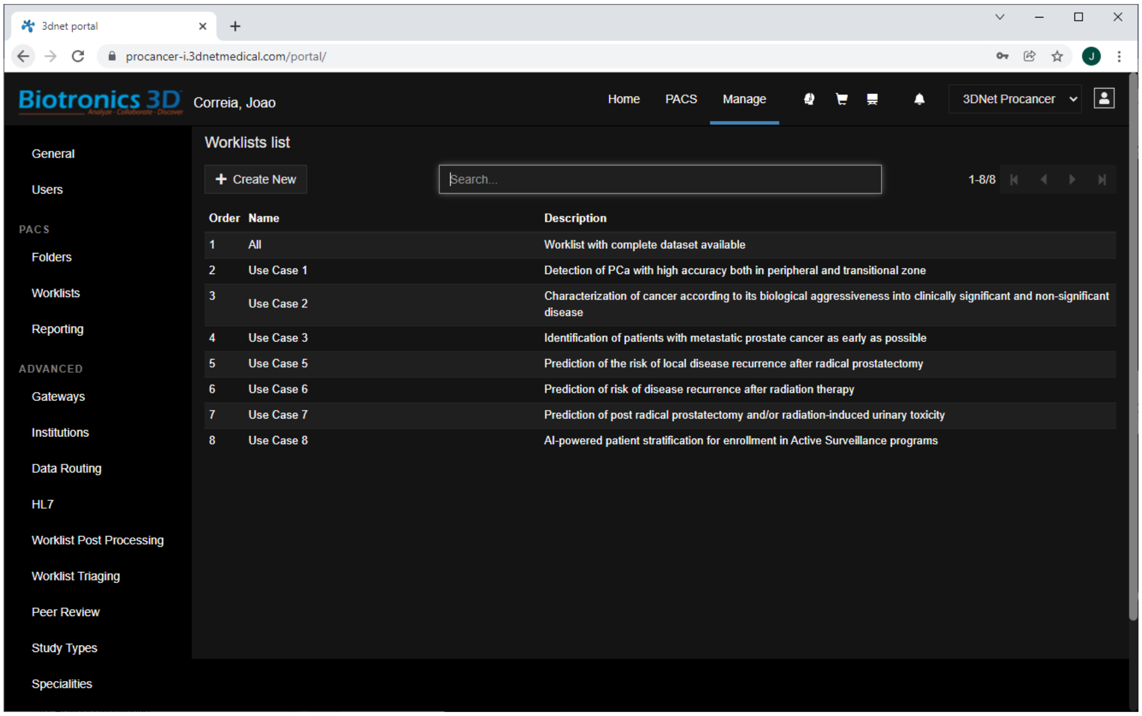
User - [redacted] (OrgAdministrators)

☐ Role based default permissions

Worklists Permissions: All, Use Case 1 ×, Use Case 2 ×, Use Case 3 ×, Use Case 5 ×, Use Case 6 ×, Use Case 7 ×, Use Case 8 ×

Name										
@ProstateNet	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
All	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
FCHAMPALIMAUD	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
FPO	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
GAONA	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
HACETTEPE	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
HULAFE	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
IDIBGI	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑
IPC	☑	☑								
JCC	☑	☑								
NCI	☑	☑	☑	☑	☑	☑	☑	☑	☑	☑

Figure 38. Management of a user’s access rights

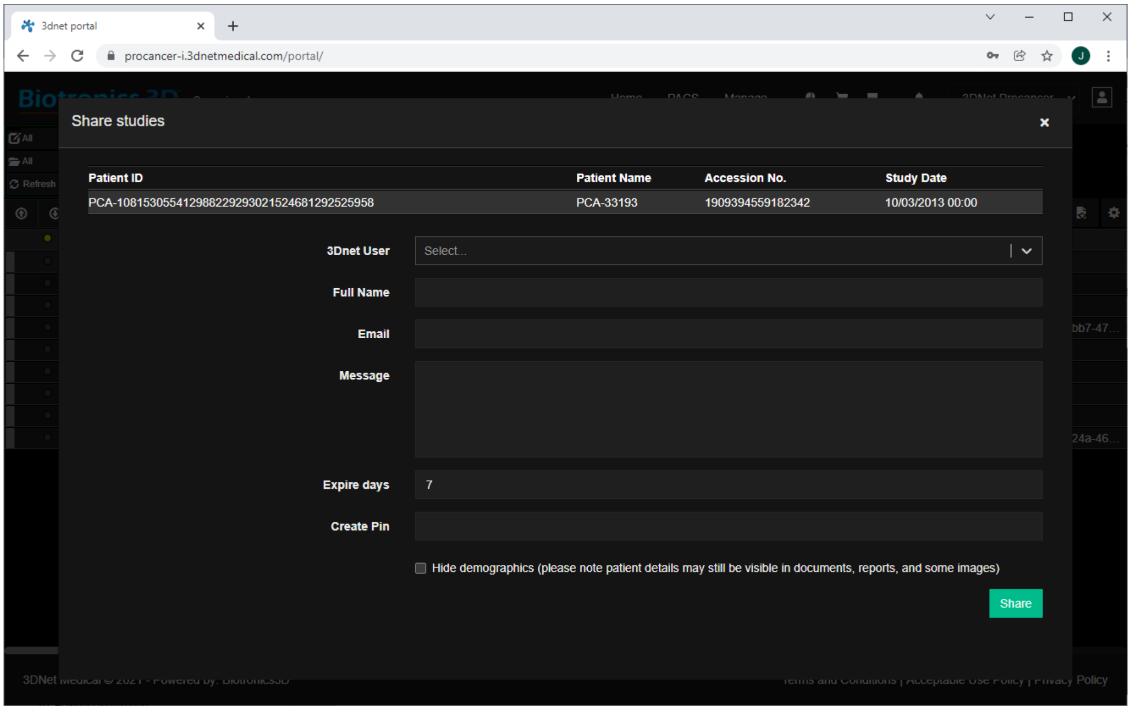


Worklists list

Order	Name	Description
1	All	Worklist with complete dataset available
2	Use Case 1	Detection of PCa with high accuracy both in peripheral and transitional zone
3	Use Case 2	Characterization of cancer according to its biological aggressiveness into clinically significant and non-significant disease
4	Use Case 3	Identification of patients with metastatic prostate cancer as early as possible
5	Use Case 5	Prediction of the risk of local disease recurrence after radical prostatectomy
6	Use Case 6	Prediction of risk of disease recurrence after radiation therapy
7	Use Case 7	Prediction of post radical prostatectomy and/or radiation-induced urinary toxicity
8	Use Case 8	AI-powered patient stratification for enrollment in Active Surveillance programs

Figure 39. Management of worklists on an organisation

Users with an OrgAdmin role can manage access to folders and worklists for users in that organisation. Users with roles of Reporting Physicians and Technologists can share studies that are available to them on folders with permissions, as shown in Figure 40.



Share studies

Patient ID	Patient Name	Accession No.	Study Date
PCA-108153055412988229293021524681292525958	PCA-33193	1909394559182342	10/03/2013 00:00

3Dnet User: Select...

Full Name:

Email:

Message:

Expire days: 7

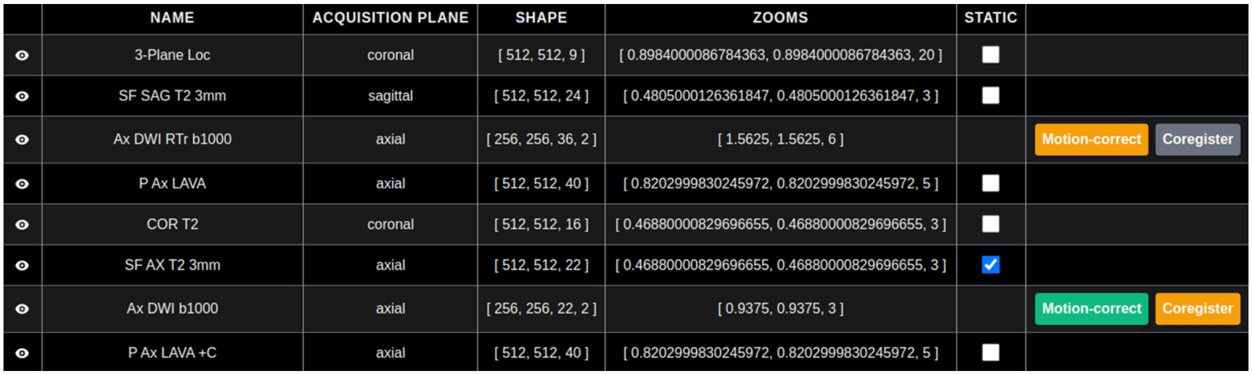
Create Pin:

□ Hide demographics (please note patient details may still be visible in documents, reports, and some images)

Figure 40. Sharing studies by users with access rights

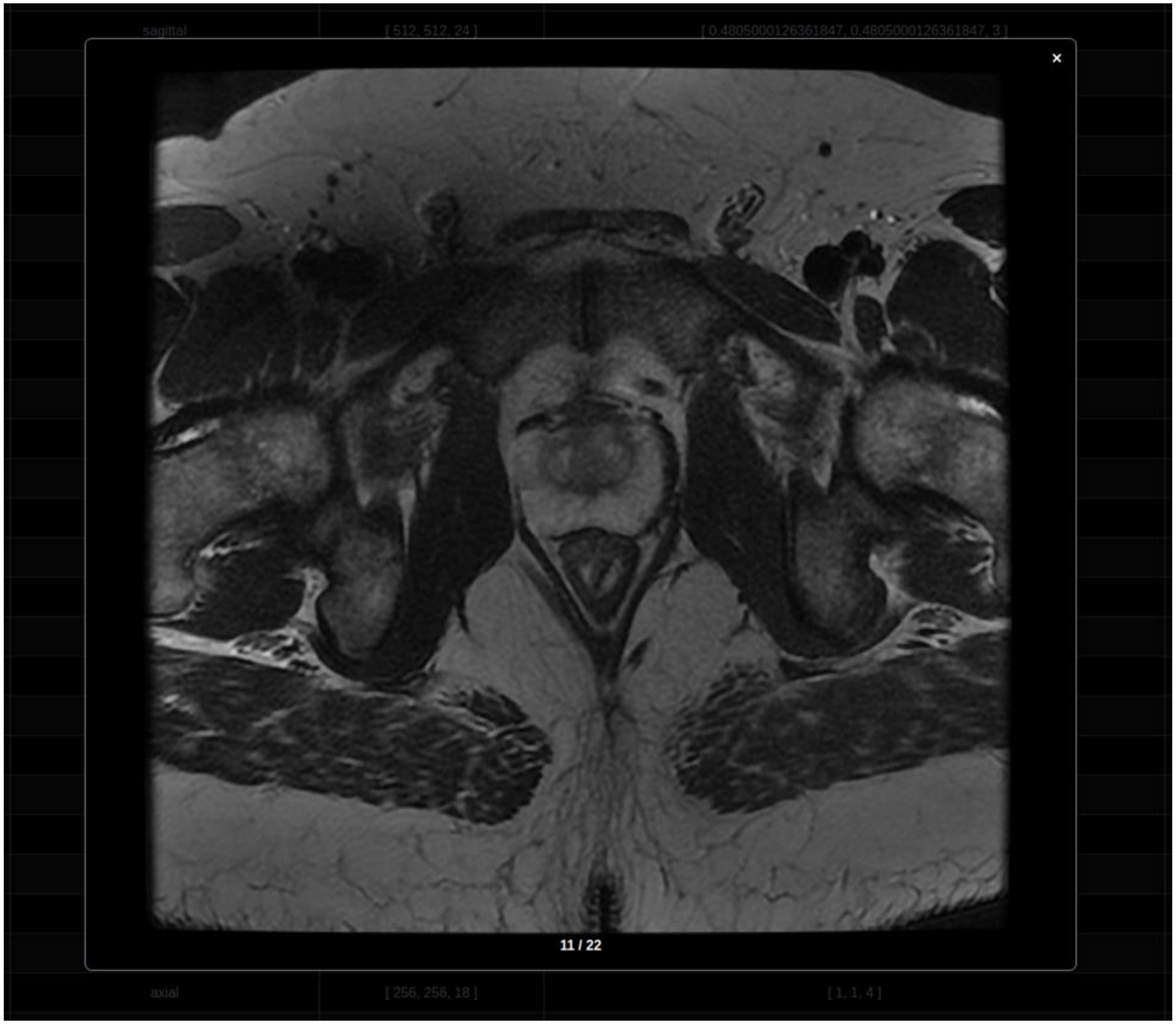
5.6 Data Curation

The image curation tool guides the user through a set of steps for the curation: motion-correction, co-registration, and quality check, and final approval and storage of the derived images or their rejection, after the manual inspection of the results. On the ProCAncer-I platform the curation tool is instantiated at the Cloud Staging Area with a Study Instance UID. The frontend component is instantiated using the Study Instance UID, which is also passed on to the backend application in order to retrieve the corresponding DICOM instances in order to be subsequently displayed to the user. Initially, a study can be either identified by its Study Instance UID and downloaded via a DICOMweb RESTful API or uploaded as a ZIP/TAR file to the server. In the pre-processing stage, the curation tool attempts to identify all included studies, in case multiple studies are present, especially in a manual upload, and group the DICOM series per study. Afterwards, certain public DICOM tags, such as the Series Instance UID and Series Description, alongside additional metadata, like the image's shape, zooms, and plane of acquisition, are extracted to serve as indexing keys, user-facing information, and inputs to upcoming processing steps. During this stage, the tool also attempts to discover a default static series to be used in the image co-registration phase. The default static series must be (a) 3D, (b) T2w (based on the series description and/or protocol name), (c) axially acquired (based on the image's zooms extracted from the affine matrix), and (d) not the result of a curation function. At the end of this pre-processing phase, a series of asynchronous tasks are scheduled to preemptively compute curation outcomes with default parameters. Figure 41 below shows the list of series found in a study with accompanying metadata, after the pre-processing phase has successfully finished. The "eye" icon on the left-hand side of each row can be used to view the corresponding image in a pop-up viewer, as shown in Figure 42. The *STATIC* column can be used to change the static (3D) series to be used in the image co-registration phase. The color-coded buttons on the right-hand side are meant to start and guide the user through the curation process: motion-correction, co-registration, and, finally, approval or rejection of the results.



	NAME	ACQUISITION PLANE	SHAPE	ZOOMS	STATIC	
👁	3-Plane Loc	coronal	[512, 512, 9]	[0.8984000086784363, 0.8984000086784363, 20]	□	
👁	SF SAG T2 3mm	sagittal	[512, 512, 24]	[0.4805000126361847, 0.4805000126361847, 3]	□	
👁	Ax DWI RTr b1000	axial	[256, 256, 36, 2]	[1.5625, 1.5625, 6]		Motion-correct Coregister
👁	P Ax LAVA	axial	[512, 512, 40]	[0.8202999830245972, 0.8202999830245972, 5]	□	
👁	COR T2	coronal	[512, 512, 16]	[0.46880000829696655, 0.46880000829696655, 3]	□	
👁	SF AX T2 3mm	axial	[512, 512, 22]	[0.46880000829696655, 0.46880000829696655, 3]	☑	
👁	Ax DWI b1000	axial	[256, 256, 22, 2]	[0.9375, 0.9375, 3]		Motion-correct Coregister
👁	P Ax LAVA +C	axial	[512, 512, 40]	[0.8202999830245972, 0.8202999830245972, 5]	□	

Figure 41. Curation Tool UI - Series List



sagittal	[512, 512, 24]	[0.4805000126361847, 0.4805000126361847, 3]
axial	[256, 256, 18]	[1, 1, 4]

Figure 42. Curation Tool UI - Image Viewer

The color-coding, as shown in Figure 43, indicates the state of the entire curation workflow per series. A greyed-out button, e.g. *Coregister,* in the case below, cannot be clicked, unless the previous stage is green. A green button indicates a complete step, which may be re-executed. A yellow button indicates a pending step. All pending steps must be completed in order for the curation of the particular series to be completed.

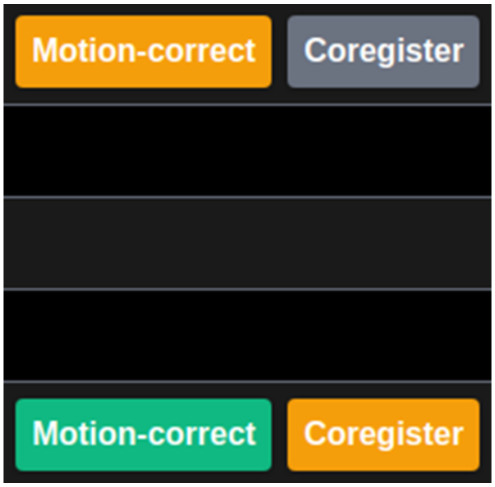
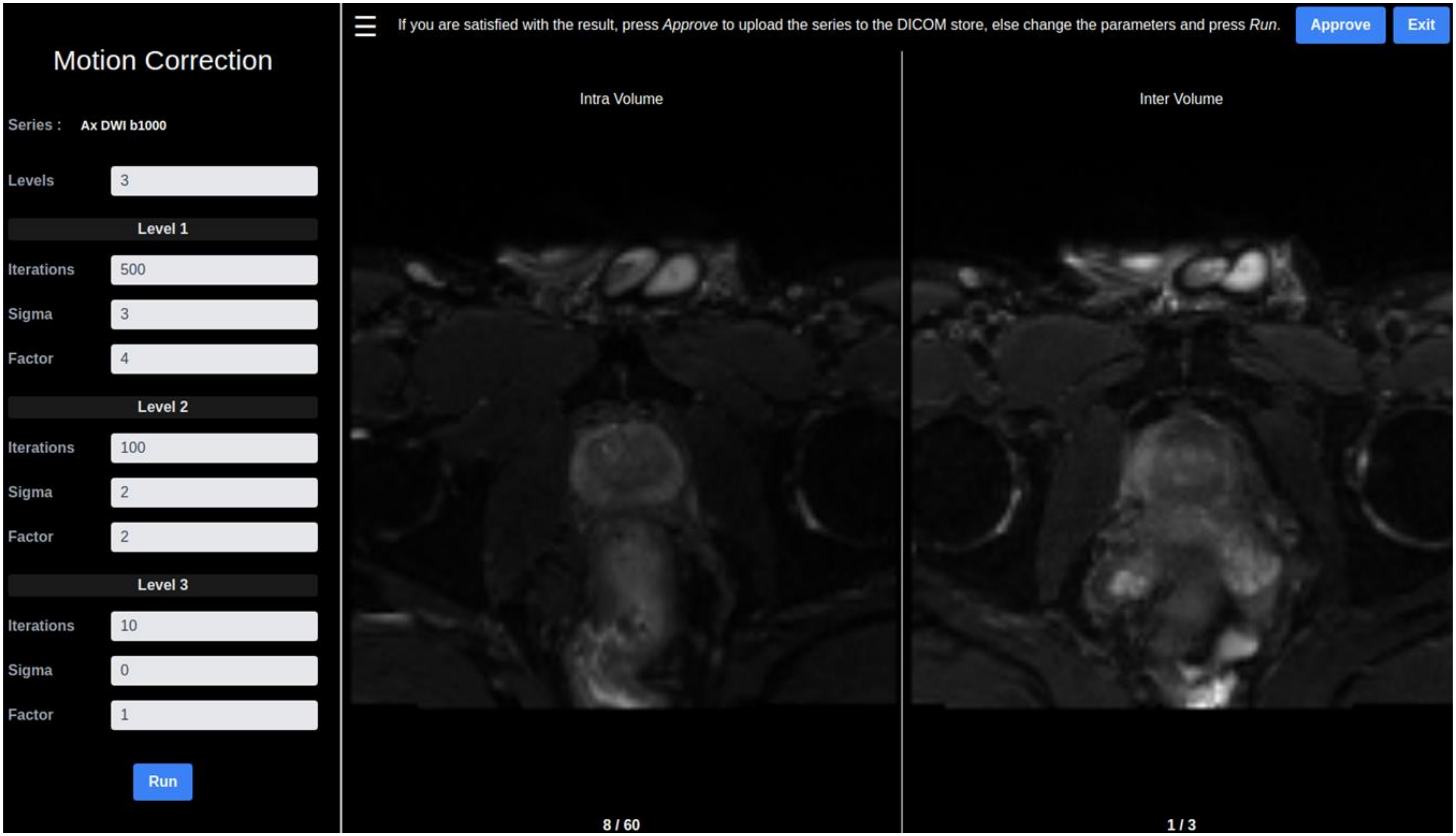


Figure 43. Curation Tool UI - State Indicators

As such, the curation of a series is a guided process. The UI's (outer) series list, which is implemented as a wrapper around the (inner) motion-correction and co-registration applications, enforces the proper order of execution of all curation steps and walks the user through the entire curation process for 4D (DWI and DCE) series step by step.

The motion-correction application performs inter-volume motion-correction of a DWI or DCE series by computing an affine transformation to register two 3D volumes. The first volume is automatically selected as the reference volume. The selection of the reference volume, as well as the exclusion of slices with high levels of motion, could be made configurable by the user in future versions. The motion-corrected series can be concurrently reviewed for intra- and inter-volume motion in two side-by-side viewers. The user may scroll through all the slices of all volumes and the middle slice per volume, respectively, in order to identify excess motion. For instance, in Figure 44, on the left-hand side of the split screen the user can scroll through all total 60 slices of all volumes (20 slices x 3 volumes). On the right-hand side, the user can scroll through the middle slice of each of the 3 volumes. Figure 44 shows the results for slice 8/60 (8th slice of the 1st volume) and volume 1/3, respectively. If the result is satisfactory, the *Approve* button on the top-right of the screen, as shown in Figure 44, accepts the result by sending it to the Cloud Staging Area and advances the user to the co-registration application taking as input the result of the motion-correction stage. On the other hand, if the result is unsatisfactory, the registration hyperparameters on the left-hand side may be refined in order to re-motion-correct the image.



Motion Correction

Series : Ax DWI b1000

Levels: 3

Level 1

Iterations: 500

Sigma: 3

Factor: 4

Level 2

Iterations: 100

Sigma: 2

Factor: 2

Level 3

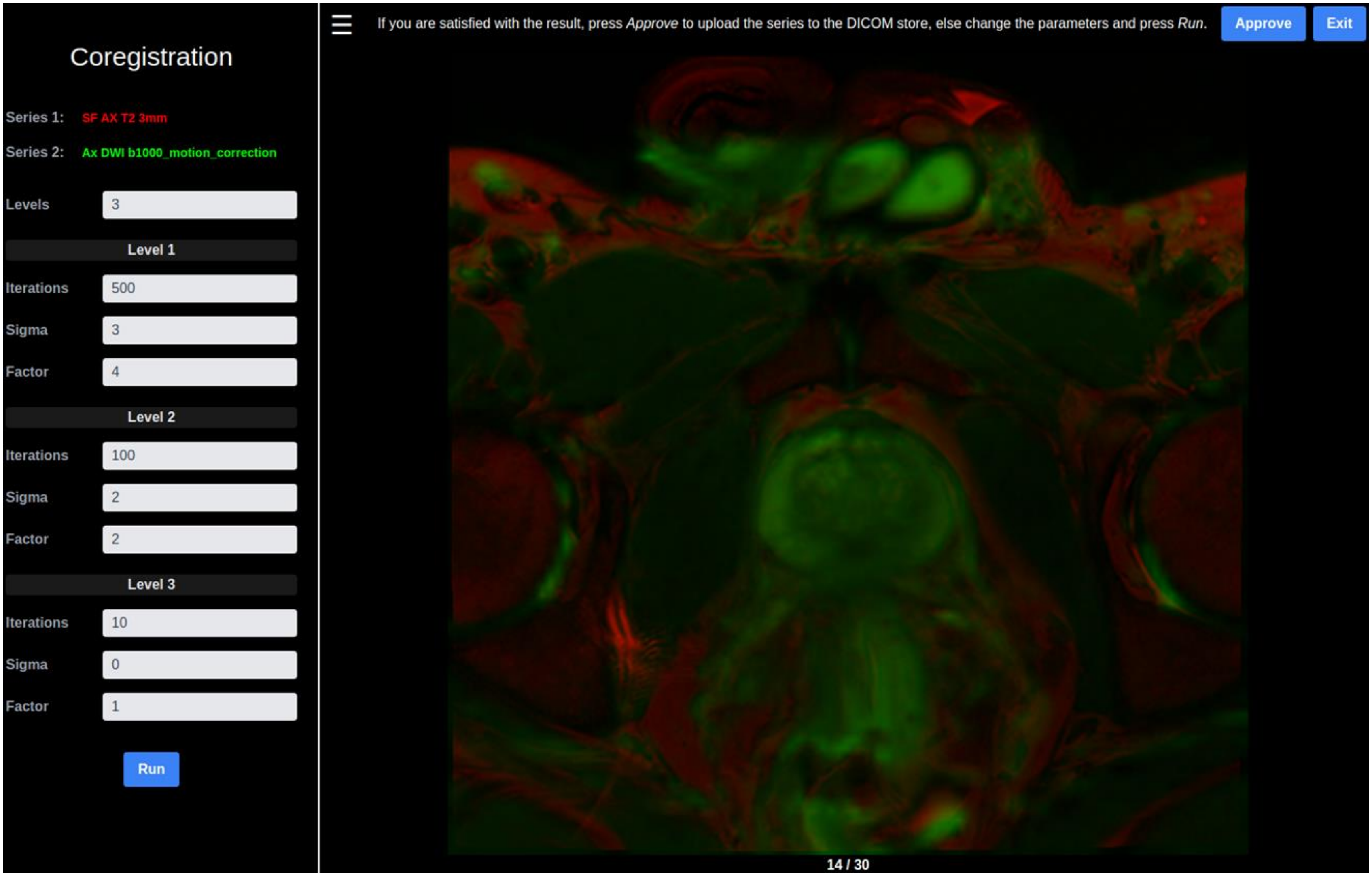
Iterations: 10

Sigma: 0

Factor: 1

If you are satisfied with the result, press *Approve* to upload the series to the DICOM store, else change the parameters and press *Run*.

Figure 44. Curation Tool UI - Motion Correction Application



Coregistration

Series 1: SF AX T2 3mm

Series 2: Ax DWI b1000_motion_correction

Levels: 3

Level 1

Iterations: 500

Sigma: 3

Factor: 4

Level 2

Iterations: 100

Sigma: 2

Factor: 2

Level 3

Iterations: 10

Sigma: 0

Factor: 1

If you are satisfied with the result, press *Approve* to upload the series to the DICOM store, else change the parameters and press *Run*.

Figure 45. Curation Tool UI - Co-registration Application

The co-registration application co-registers the motion-corrected series to a T2w image. The moving (motion-corrected) and static (T2) series are also color-coded in green and red, respectively. In Figure 45, the 1st volume of the previously motion-corrected series has been laid over the T2w axial series and the result for slice 14/30 is illustrated. If the result of the co-registration is unsatisfactory, the registration hyperparameters may be refined in order to re-run the process. If the result is acceptable, by clicking the *Approve* button the user accepts the new, co-registered image

Future versions of the motion-correction and co-registration UIs could also offer users the ability to modify display settings, such as colouring and transparency, in order to improve inspection.

All computed results are cached on the local filesystem of the server and they are immediately returned from the cache if re-requested. Additionally, the caching mechanism exploits a per-session unique key in order to isolate processing per user. This means that multiple users may start working on different studies without cross-user interference. Additionally, multiple users may curate the same study, produce the same or different results, based on the registration hyperparameters used for the motion-correction and co-registration algorithms, and store all newly derived images on the DICOM store.

During curation, a DICOM series may be motion-corrected once. If the same image is motion-corrected again with the same set of registration parameters, the previously cached result will be immediately returned. If a different set of registration parameters is used, a new image will be produced and the result will overwrite the previous one. The same applies for image co-registration. However, in this case, if the same parameters are used, but a different static series is selected, a new image will be created that will not overwrite any previous result. Lastly, if a series has been fully curated and the user attempts to re-run motion-correction with a different set of parameters, all series derived from the previously motion-corrected series will be invalidated and the state of curation for the specific series will reset, since the previously motion-corrected series will get overwritten. Note, however, that already approved results, which have been stored on the Staging Area or DICOM Store, will remain unaffected.

After curation of the study has finished, each new image is uploaded to the Cloud Staging Area as a new DICOM series with a freshly generated Series Instance UID and any related series (of the original study) added to the Referenced Series Sequence public DICOM tag and curation-related information is added to the Metadata Catalogue. If the study is selected for curation again, the already approved and stored outputs are not modified or deleted. Curation results are stored in an append-only fashion. Such an approach makes it easier for multiple users to curate the same study and produce different results.

As already shown on the left-hand side in Figure 44 and Figure 45, the inputs for configuring the motion-correction and co-registration algorithms are the same as shown. The underlying core registration algorithm uses a combination of transformation steps: translation, rigid, and affine to compute an affine transformation to register two 3D volumes via a similarity metric. Each step is performed at multiple levels to align two volumes using a multi-resolution strategy. The number of levels specify the number of resolutions. For each level the number of iterations (to solve the optimization problem) is specified along with a smoothing (Guassian kernel sigma) and a scaling (downsampling) factor. It is recommended to exploit higher smoothing and more downsampling for more iterations in early levels in order to achieve an early, good-enough transformation before more fine-grained transformations towards the end. The prefilled parameters should work for most cases. Finally, the curation tool exposes the following API endpoints:

Table 22. *Curation Tool API Endpoints*

Method	Path	Query Parameters	Request Body
GET	./studies		
	Return all uploaded studies		
POST	./studies		study_instance_uid data[23]
	Upload a new study archive or download a study from the DICOM store		
GET	./studies/<study_uid>		
	Return metadata of all the study's series		
POST	./studies/<study_uid>/approve		
	Approve all of the study's derived series		
GET	./studies/<study_uid>/series/<series_uid>		
	Return the series' metadata		
GET	./studies/<study_uid>/series/<series_uid>/data	middle_slc plane out_shape[24]	
	Return a spritesheet of the series' (middle) slices across the specified plane in the		

[23] The uploaded file is expected to be under the data key in a multipart/form-data request.

[24] The middle_slc defaults to False. The plane defaults to axial. The out_shape defaults to None.

Method	Path	Query Parameters	Request Body
	given shape[25] [4]		
POST	./studies/<study_uid>/series/<series_uid>/motion-correct		parameters [26]
	Apply motion-correction to the specified series		
POST	./studies/<study_uid>/series/<series_uid>/coregister	static_series_uid	parameters [25]
	Coregister the specified (moving) series		
POST	./studies/<study_uid>/series/<series_uid>/approve		
	Approve the specified, derived series		

The above API endpoints allow the upload of new studies, retrieval of studies' and series' metadata from the curation tool's cache, the execution of curation functions, such as motion-correction, and the generation of sprites to view in the pop-up viewers.

5.7 Data Annotation

For the annotation of the DICOM studies incorporated in the ProCancer-I platform, an annotation tool environment has been developed and is ready to be integrated with the rest of components of the ProCancer-I system.

This tool has been designed to follow a DICOM in – DICOM out approach. Therefore, both the inputs and outputs of this environment will be DICOM files and must follow the DICOM standard (NEMA PS3 / ISO-12052 Digital Imaging and Communications in Medicine (DICOM) Standard).

As a main feature, it provides the ability to draw regions of interest (ROI) on the medical image by using a brush. The brush allows the user to easily segment regions of interest in the image by marking the desired pixels (Figure 46).

In addition, there are other features included in the annotation tool:

- DICOM series arrangement and visualization and exploration through sliders.
- DICOM header visualization and exploration.
- Add notes to the study.
- Change zoom and pan.
- Change image contrast.

[25] Returns an image/png response. All other API endpoints return a JSON payload

[26] The parameters must follow the curation metadata's parameters in Appendix B.

- Change image colormap.
- MPR (Multi Planar Reconstruction) visualization.
- Split visualization.
- ROIs by label organization.
- ROIs removal.
- Undo/Redo.
- ROIs quantification.

More information about these functionalities can be found in deliverable D4.2. Data annotation, curation and upload tools.

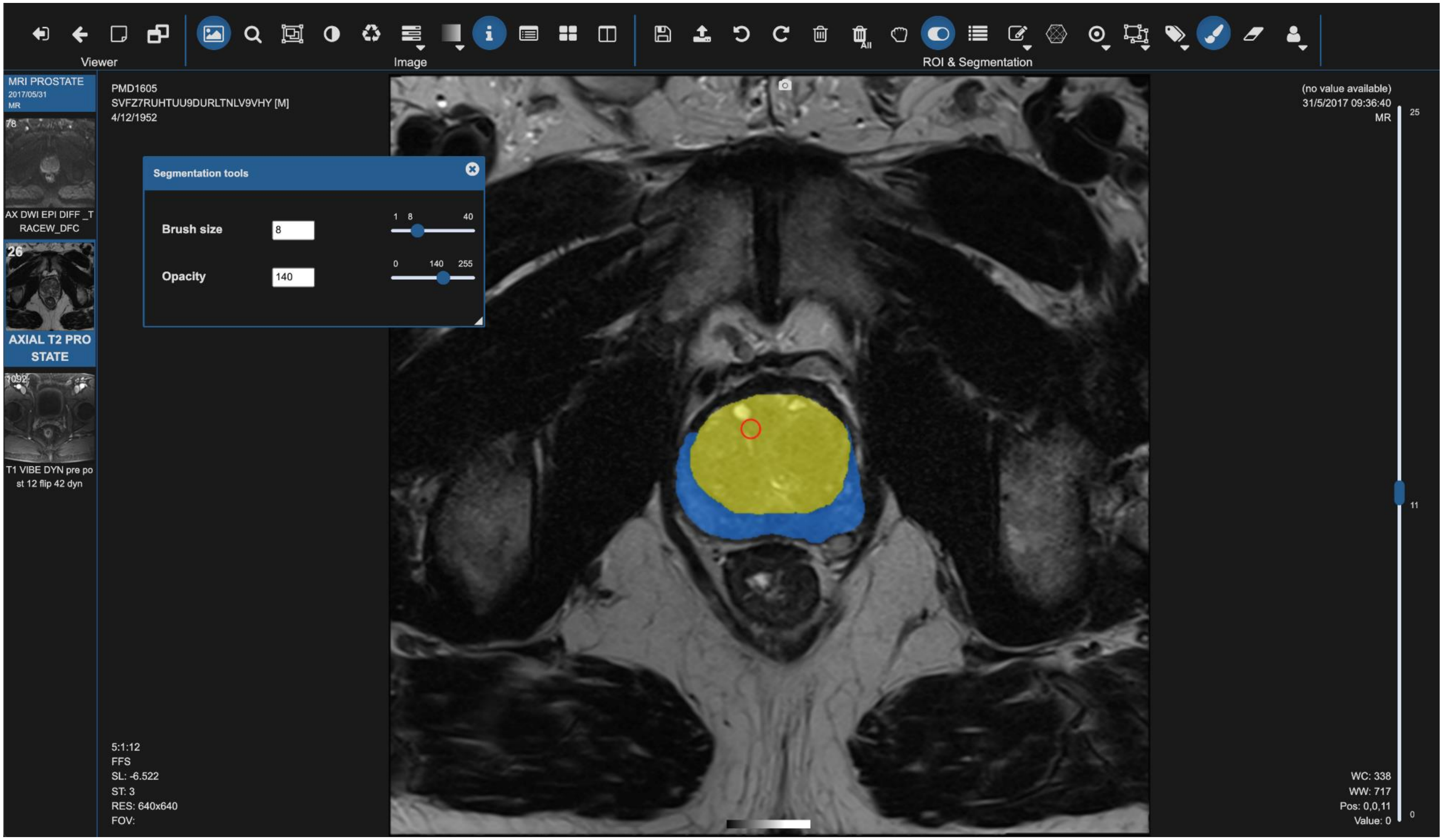
Annotations made with the brush can be marked to be stored in the ProCAncer-I database. These segmentation files are stored as DICOM Seg objects.

(http://dicom.nema.org/medical/Dicom/2018d/output/chtml/part03/sect_C.8.20.html#table_C.8.20-1).

These are DICOM files intended to store information about labeled voxels. Among its main benefits are:

- Size efficiency with multi-frame storage and bit encoding.
- Structured terminology for encoding semantics.
- Binary and fractional segmentation (e.g., probability maps)
- Encoding of the presentation (e.g., color and window level)
- Multiple voxel occupancy (i.e., a single voxel of the source image can have multiple labels).
- Segmentation details: algorithm name, type, version, parameters
- Anatomic region details
- Content creator details: name, id, institution
- Integration with DICOM object family:
  - Addition as a new studio series with unequivocal relationship with patient and study.
  - Reference with source image.
  - It can be referenced from the measurement documents (DICOM SR)
- The annotation tool also allows the import of DICOM Seg files to be displayed on top of the original image. This allows the user to load and edit annotations made in 3rd party software to be integrated into the project database.
- Apart from the segmentation file other metadata can be generated and stored in the Meta-Data Catalogue.
- This annotation tool is a web application built with JavaScript and HTML5 technologies. Through REST requests it communicates with the ProCancer-I database in order to retrieve and store the DICOM data. This application has been embedded in a Docker container in order to be deployed together with the other components of the platform.

Docker is a tool that allows to easily create, deploy and run applications using containers, so that each container wraps the main code together with the required libraries, dependencies, and any other execution requirement. In this way, Docker containers run as isolated applications regardless of the execution environment (Operating System or OS, hardware, etc.). This application is accessible via URL to be opened by other components of the platform.



PMD1605
SVFZ7RUHTUU9DURLTNLV9VHY [M]
4/12/1952

Segmentation tools

Brush size 8 (1 – 40)

Opacity 140 (0 – 255)

5:1:12
FFS
SL: -6.522
ST: 3
RES: 640x640
FOV:

(no value available)
31/5/2017 09:36:40
MR

WC: 338
WW: 717
Pos: 0,0,11
Value: 0

Figure 46. Prostate segmentation using the annotation environment.

6. AI Models Framework

DevOps stands for a set of practices and tools for developing, testing, deploying, and operating large-scale software systems, generally organized in one solution, or framework. With DevOps, development cycles became shorter, deployment velocity increased, and system releases became auditable and dependable.

Machine learning pipelines require solutions like DevOps, or even smarter, since ML systems are experimental in nature, and have more components that are significantly more complex to build and operate. Thus, recently, there have been developed many proprietary and open source AI models frameworks; also the concept of MLOps arose.

MLOps is a set of "best practices" for managing the entire Machine Learning lifecycle, requiring awareness of common issues, and efficient communication between data scientists and operations professionals. This kind of approach is essential when you want to build, train, and validate reproducible ML pipelines, or to increase the process accountability and trustworthiness.

MLOps methodology includes a process for streamlining model training, packaging, validation, deployment, and monitoring.

Experiment tracking is a very important part (or process) of MLOps, which is focused on collecting, organizing, and tracking model training information across multiple runs with different configurations (hyperparameters, model size, data splits, parameters, and so on).

Another very important part of MLOps is Model management, and, in particular, model monitoring. A logical, easy-to-follow policy for Model management is essential to ensure that ML models are consistent, and all requirements are met at scale.

In general, the phases of MLOps are:

- Data gathering
- Data analysis
- Data transformation/preparation
- Model training & development
- Model validation
- Model serving
- Model monitoring
- Model re-training

While the key ingredients of MLOps are:

- Model metadata storage and management
- Data and pipeline versioning
- Hyperparameter tuning
- Run orchestration and workflow pipelines
- Model deployment and serving
- Production model monitoring

6.1 Experiment Tracking

Developing AI models often requires evaluating a large number of features, parameters and/or datasets and iterating over several experiments. In order to obtain the best model that will be moved to production, tracking all experiments for comparability and reproducibility is critical. Therefore, experiment tracking focuses on the iterative model development phase where different things are tested until we get the desired model performance. It can be divided in three phases:

Model training: It involves ingesting data, engineering new features and monitoring the training process. Training can be done over several experiments. Each of the experiments is logged in detail and once the process is complete, the experiment that resulted in the best performance can be reproduced.

Model evaluation: At this stage the model is tested with data that it has never seen before, in readiness for deployment. The test data features should be similar to the ones used in training. Otherwise evaluation will fail. If the performance is dismal then retraining the model can be done after adjusting the features and/or the data.

Model registry: By logging the models themselves, one can immediately pick the model that resulted in the best performance and use it for serving. The model registry can be used in the retraining phase in model management as well. Once a new model is available, the model registry has to be updated. The model registry contains the model metadata which are important for (i) managing the models, (ii) complying with regulatory frameworks and (iii) knowing if the model is running in production and at which end point.

6.2 AI Model Monitoring

Once the best performing model is registered and served, it needs to be monitored so that in the event of a performance drop, the necessary measures can be taken. When the conditions dictate so, retraining and remodelling can be executed.

Some examples where AI model monitoring is required are:

- **monitor model performance in production**: assess how accurate the predictions of the model are. If the model performance decays over time (model drift) then you might need to retrain it.
- **monitor model input/output distribution**: assess whether the distribution of input data and features that go into the model have changed (data and concept drift), such as the predicted class distribution changed over time.
- **monitor model training and re-training**: evaluate learning curves, trained model predictions distribution, or confusion matrix during training and re-training.
- **monitor model evaluation and testing**: log metrics, charts, prediction, and other metadata for automated evaluation or testing pipelines
- **monitor hardware metrics**: see how much CPU/GPU or Memory the models use during training and inference.

6.3 Experiment Tracking ProstateNet Module

In order to support the experiment tracking and AI model monitoring, certain services should exist. At the moment many proprietary and open source frameworks have been developed. Well known frameworks are: FBLearner Flow[27] which is a machine learning platform capable of easily reusing algorithms in different products, scaling to many simultaneous custom experiments, and managing experiments with ease; Michelangelo[28] which enables internal teams to seamlessly build, deploy, and operate machine learning solutions; TensorFlow Extended (TFX)[29] which is an end-to-end platform for deploying production ML pipelines. Currently, among the existing solutions, KubeFlow and MLflow earn a special mention.

KubeFlow is an open source cloud-native platform for machine learning operations. Kubeflow started as a project to make deployments of ML workflows on Kubernetes simple, portable and scalable, integrating various frameworks. It is supported by a large community of users and contributors (GitHub), organised into working groups. It includes tools for creating and managing pipelines (using Docker containers), for ML training (with a custom TensorFlow job operator), for the management of Jupyter notebooks (e.g. building notebook containers, or pods directly in clusters), and ML models deployment (through add-ons).

Both Kubeflow and MLflow use cutting-edge technologies and tools to create a collaborative platform for model development; but Kubeflow, with respect to MLflow, shows an overly complexity, being rather demanding to set up and maintain: e.g. the exploratory data analysis (EDA) can be performed in both Kubeflow and MLflow, but using Kubeflow requires higher technical know-how. Also, Kubeflow is, at its core, a container orchestration system; while MLflow is a Python program for tracking experiments and versioning models. Hence, for example, when you train a model in Kubeflow, everything happens within the system, while with MLflow, the actual training happens wherever you choose to run it, and the MLflow service listens in on parameters and metrics. As the ProCAncer-I platform will be used by a large community of data scientists and domain experts, it will be developed and deployed exploiting the MLflow framework.

MLflow[30] provides solutions for managing the ML process and deployment and it can be incorporated into any programming ecosystem. It can do experimentation, reproducibility, deployment, or be a central model registry. It consists of:

i) Tracking,
ii) Projects,
iii) Models and
iv) Model Registry.

[27] https://engineering.fb.com/2016/05/09/core-data/introducing-fblearner-flow-facebook-s-ai-backbone/

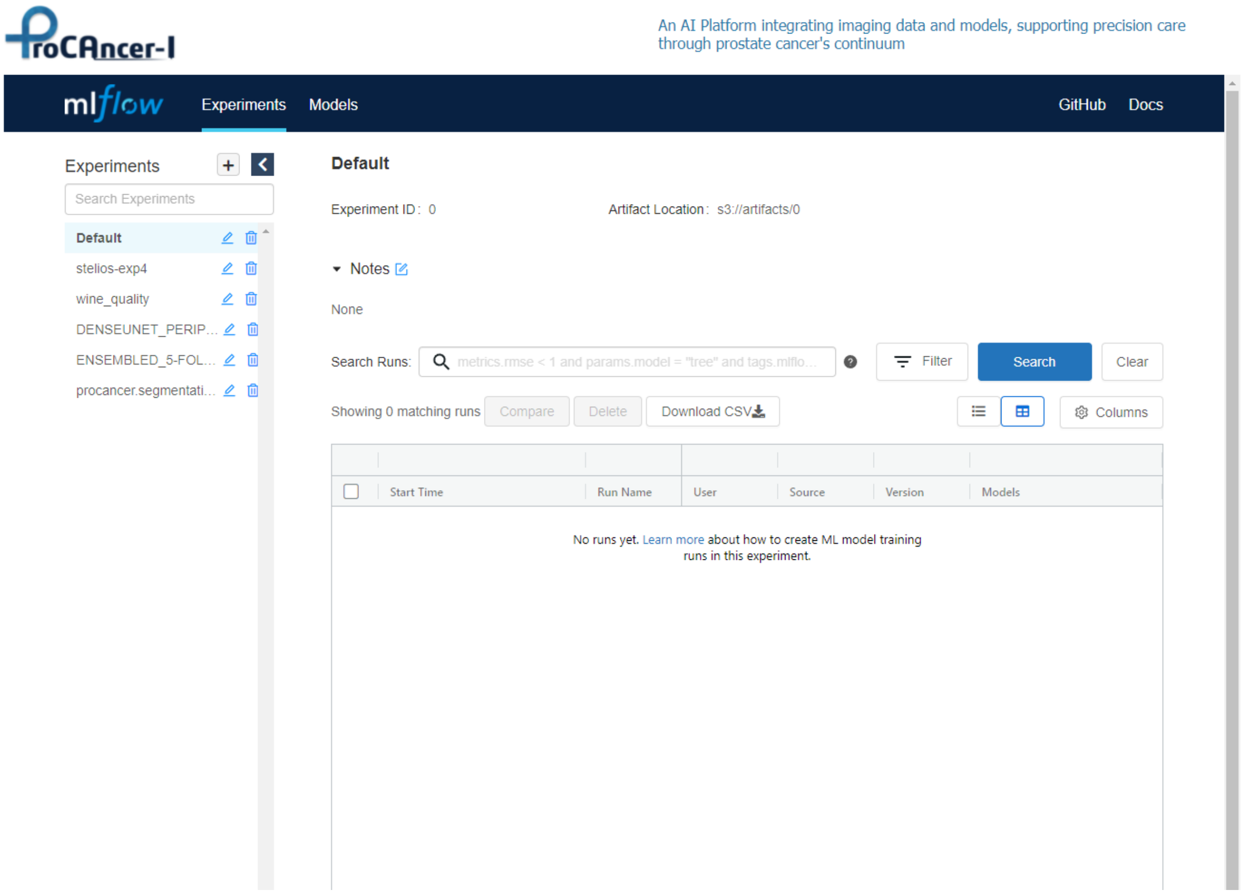
[28] https://eng.uber.com/michelangelo-machine-learning-platform/

[29] https://www.tensorflow.org/tfx

[30] https://mlflow.org/

The concept of each component respectively are: i) Maintains logs of the running experiments of the model, with all the changes implemented in each version. Also, it provides visualization of these changes. MLflow tracking is an API and UI uses Python, REST, R API, and Java API APIs. ii) Describes and organizes the code, based on conventions, ensures reusability. iii) Saves a model based on the conventional format of one or more development environments (e.g. anaconda, sklearn, aws, etc.). iv) It provides model lineage, model versioning, stage transitions, and annotations. ML flow provides flexibility, easy setup, and can be used both in local or a remote host. It runs on command, it saves the whole working environment, thus ensuring reproducibility of the code (same results), provides some standardization in how to package and deploy a model and provides an abstract description of the model, giving the ability to deploy it on different development environments. MLflow supports the export of the model into Docker containers or other commercial serving platforms. It does not support multi-users.

ProstateNet, the cloud platform of ProCAncer-I, exploits the MLFlow framework which is adapted and deployed in the development staging area (at the moment when the current deliverable is compiled) for experimenting reasons (Figure 47).



An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum

Default

Experiment ID : 0

Artifact Location : s3://artifacts/0

Notes

None

Search Runs: metrics.rmse < 1 and params.model = "tree" and tags.mlflo...

Showing 0 matching runs

□	Start Time	Run Name	User	Source	Version	Models

No runs yet. Learn more about how to create ML model training runs in this experiment.

Figure 47. MLFlow UI for the experiment tracking integrated in the ProstateNet dev staging area

In ProstateNet the MLFlow was deployed with remote Tracking Server, backend and artifact stores as depicted in the following architecture.

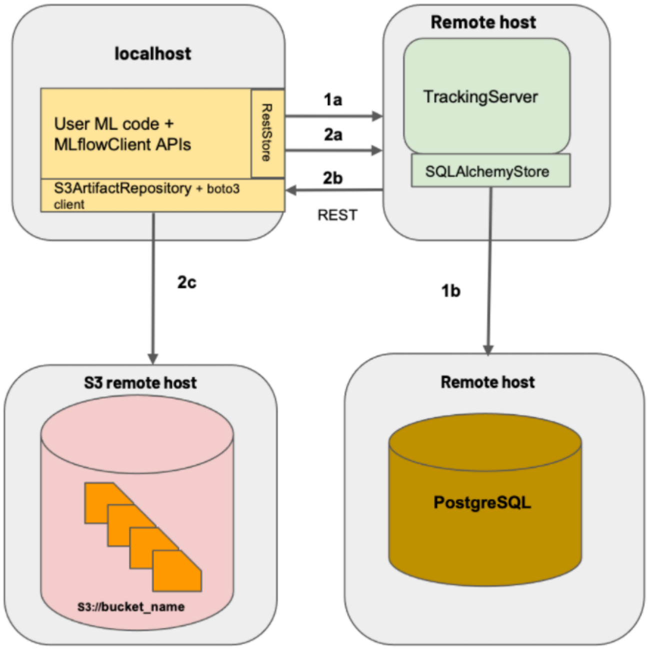


Figure 48. MLFlow integration architecture in ProstateNet dev staging area[31]

Specifically for the artifact storage an S3 server compatible alternative was deployed the MinIO[32]. The MinIO offers high-performance, S3 compatible object storage and delivers a range of use cases from AI/ML, analytics, backup/restore cases.

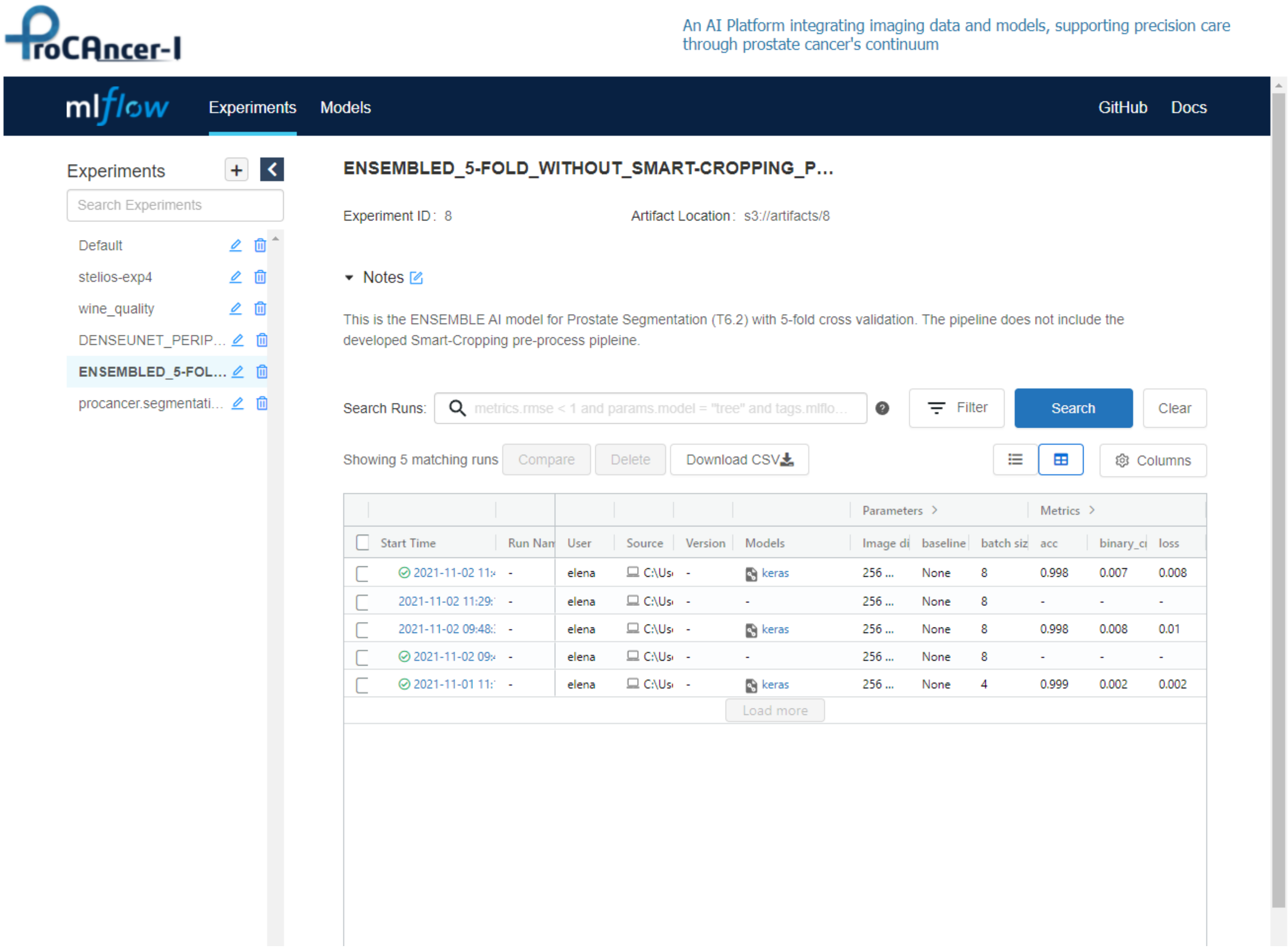
In the following section, we describe samples of the ProstateNet experiment tracking and AI model monitoring module (MLflow framework) use cases to highlight how the end-user can leverage the provided services/components.

The following figure shows the UI for the experiment tracking. Specifically, for the specific test case a Deep Learning model for the prostate segmentation is monitored. This is the ENSEMBLE AI model for Prostate Segmentation (T6.2) with 5-fold cross validation. The monitored pipeline does not include the developed Smart-Cropping pre-process. The user can record various notes for the specific experiment. Moreover, has the ability to search and filter the various runs that have been performed. The AI model list of runs provide aggregated information for the Start Time, the User, Source, Version, Type of Model (keras, scikit-learn, etc.), image information, and various performance metrics.

The user can select any of the stored results/run/model experiment and open a new UI with the run details (Figure 49, Figure 50, Figure 51 and Figure 52).

[31] https://mlflow.org/docs/latest/tracking.html

[32] https://min.io/



An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum

ENSEMBLED_5-FOLD_WITHOUT_SMART-CROPPING_P...

Experiment ID: 8 Artifact Location: s3://artifacts/8

▾ Notes

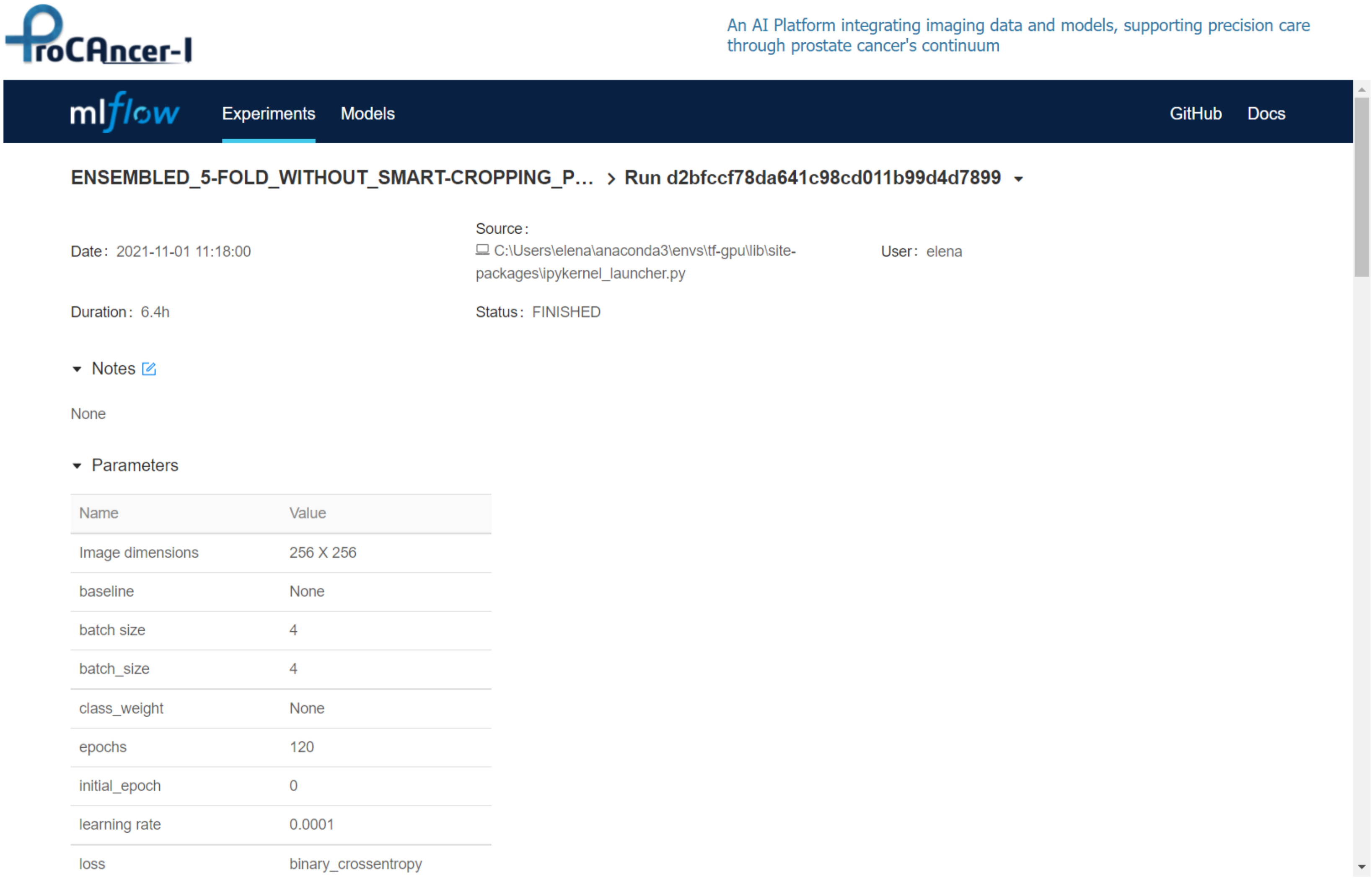
This is the ENSEMBLE AI model for Prostate Segmentation (T6.2) with 5-fold cross validation. The pipeline does not include the developed Smart-Cropping pre-process pipleine.

Search Runs: metrics.rmse < 1 and params.model = "tree" and tags.mlflo...

Showing 5 matching runs

							Parameters >			Metrics >		
□	Start Time	Run Nan	User	Source	Version	Models	Image di	baseline	batch siz	acc	binary_ci	loss
□	✓ 2021-11-02 11:	-	elena	C:\Us	-	keras	256 ...	None	8	0.998	0.007	0.008
□	2021-11-02 11:29:	-	elena	C:\Us	-	-	256 ...	None	8	-	-	-
□	2021-11-02 09:48:	-	elena	C:\Us	-	keras	256 ...	None	8	0.998	0.008	0.01
□	✓ 2021-11-02 09:	-	elena	C:\Us	-	-	256 ...	None	8	-	-	-
□	✓ 2021-11-01 11:	-	elena	C:\Us	-	keras	256 ...	None	4	0.999	0.002	0.002

Figure 49. Integrated in ProstateNet MLFlow instance for a specific experiment.


An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum
mlflow
Experiments
Models
GitHub
Docs

ENSEMBLED_5-FOLD_WITHOUT_SMART-CROPPING_P... > Run d2bfccf78da641c98cd011b99d4d7899

Date: 2021-11-01 11:18:00

Source: C:\Users\elena\anaconda3\envs\tf-gpu\lib\site-packages\ipykernel_launcher.py

User: elena

Duration: 6.4h

Status: FINISHED

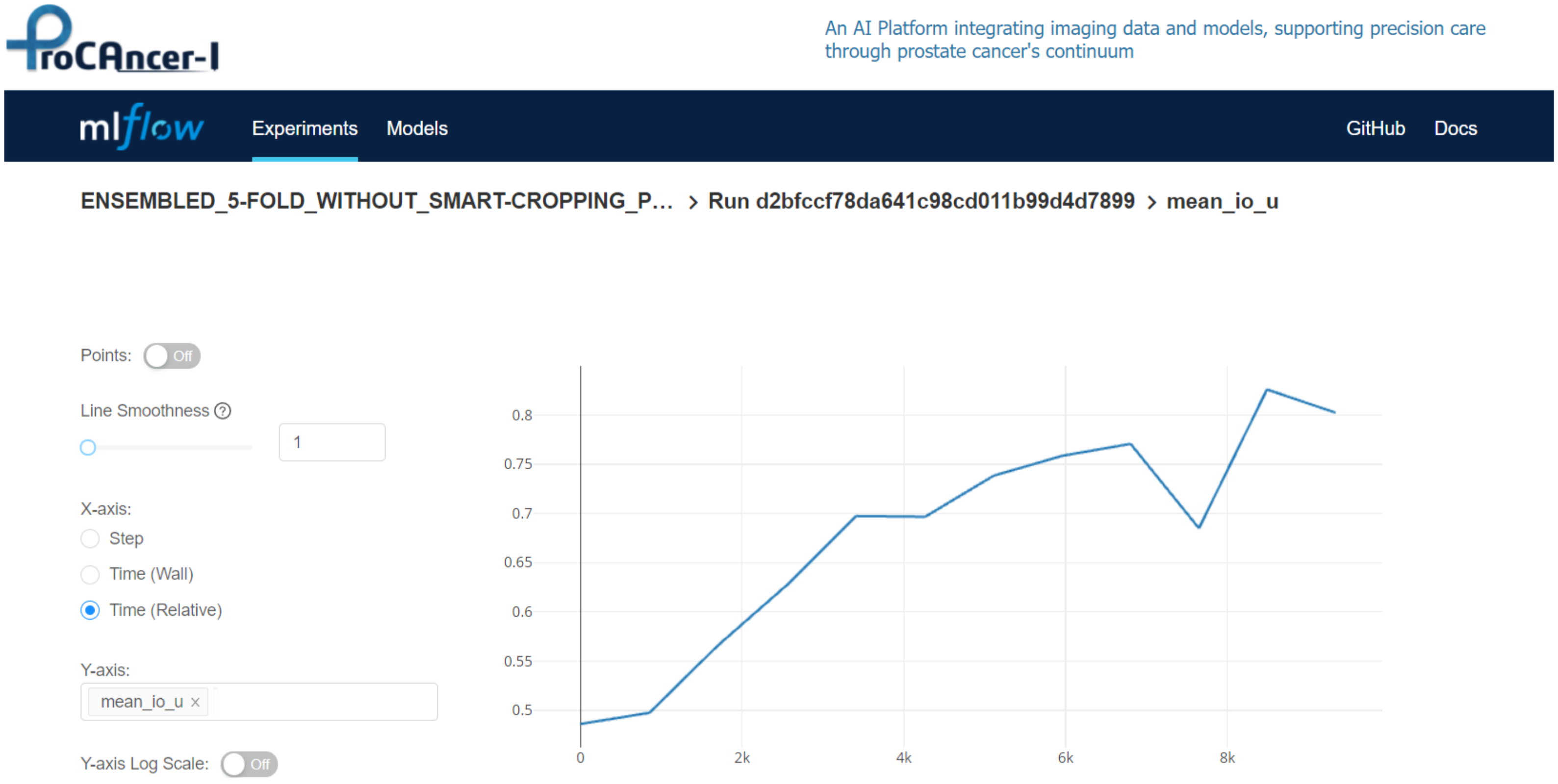
Notes

None

Parameters

Name	Value
Image dimensions	256 X 256
baseline	None
batch size	4
batch_size	4
class_weight	None
epochs	120
initial_epoch	0
learning rate	0.0001
loss	binary_crossentropy

Figure 50. Integrated in ProstateNet MLFlow instance UI with model performance metrics



An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum

ENSEMBLED_5-FOLD_WITHOUT_SMART-CROPPING_P... › Run d2bfccf78da641c98cd011b99d4d7899 › mean_io_u

Points: Off

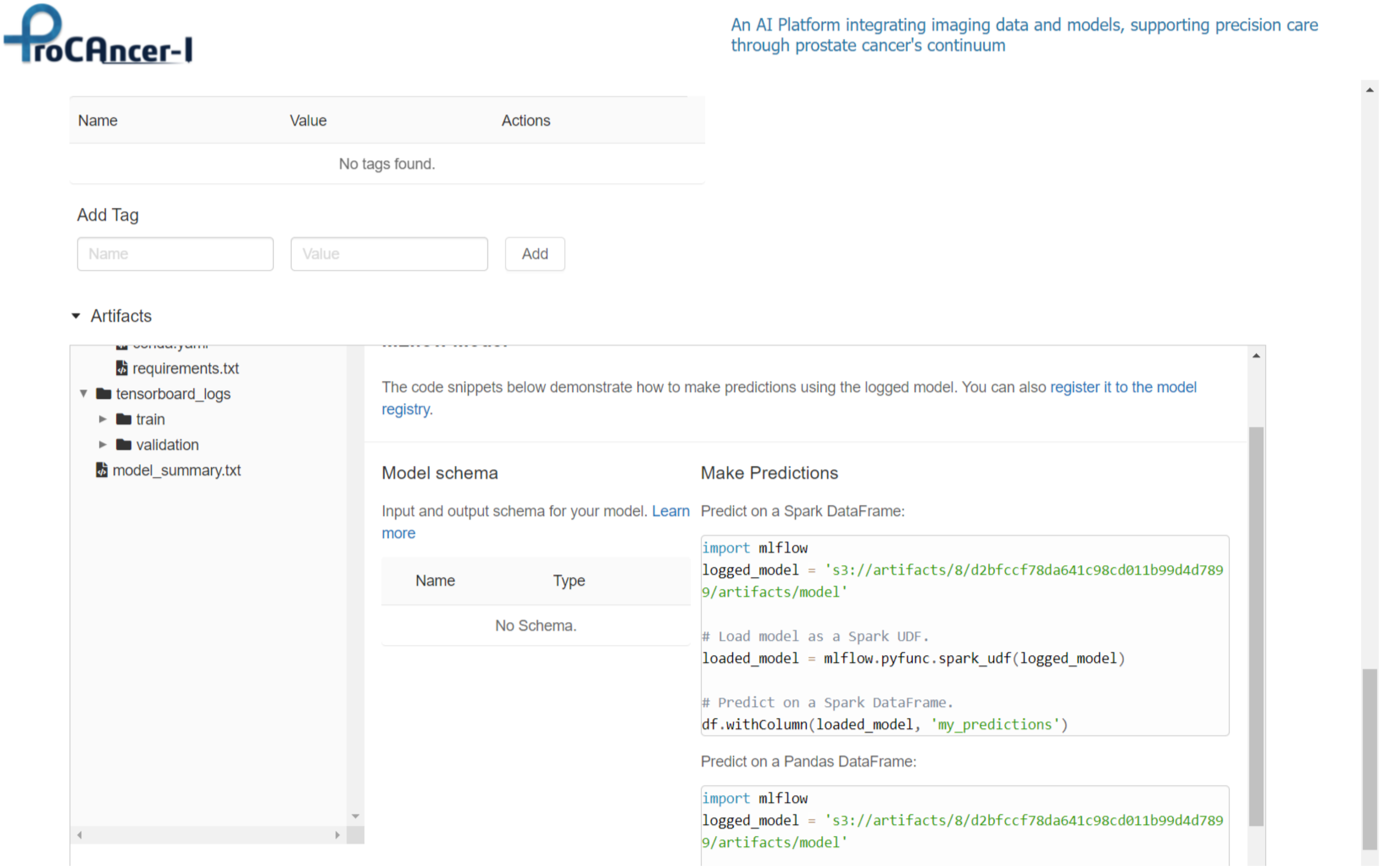
Line Smoothness: 1

X-axis:
○ Step
○ Time (Wall)
● Time (Relative)

Y-axis: mean_io_u ×

Y-axis Log Scale: Off

Figure 51. Integrated in ProstateNet MLFlow instance UI with model performance plot



An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum

Name	Value	Actions
	No tags found.	

Add Tag

▾ Artifacts

The code snippets below demonstrate how to make predictions using the logged model. You can also register it to the model registry.

Model schema

Input and output schema for your model. Learn more

Name	Type
No Schema.	

Make Predictions

Predict on a Spark DataFrame:

```
import mlflow
logged_model = 's3://artifacts/8/d2bfccf78da641c98cd011b99d4d7899/artifacts/model'

# Load model as a Spark UDF.
loaded_model = mlflow.pyfunc.spark_udf(logged_model)

# Predict on a Spark DataFrame.
df.withColumn(loaded_model, 'my_predictions')
```

Predict on a Pandas DataFrame:

```
import mlflow
logged_model = 's3://artifacts/8/d2bfccf78da641c98cd011b99d4d7899/artifacts/model'
```

Figure 52. Integrated in ProstateNet MLFlow instance UI with the model's stored artifacts

6.4 AI Model Monitoring ProstateNet Module

As previously mentioned, once the final model is selected, it is important to monitor its performance in the production environment to detect any drift may happen due to, for example, changes in the scanners or acquisition protocols. For this purpose, in the upcoming months of the project, a monitorization environment will be deployed and integrated in the ProCancer-I platform. For this purpose, it is proposed to use a solution based on Prometheus[33], an application used for event monitoring and alerting. Figure 53 shows a diagram where metrics are collected from the AI models, stored in the Prometheus server and visualized using Grafana. In addition, a Python-based statistical model will be developed to do time series analysis based on the metrics collected from the AI models and detect any drift. In the case a drift is detected, using the Prometheus Alert Manager, an alert will be generated for the users to review the status of the AI model where a drift has been detected.

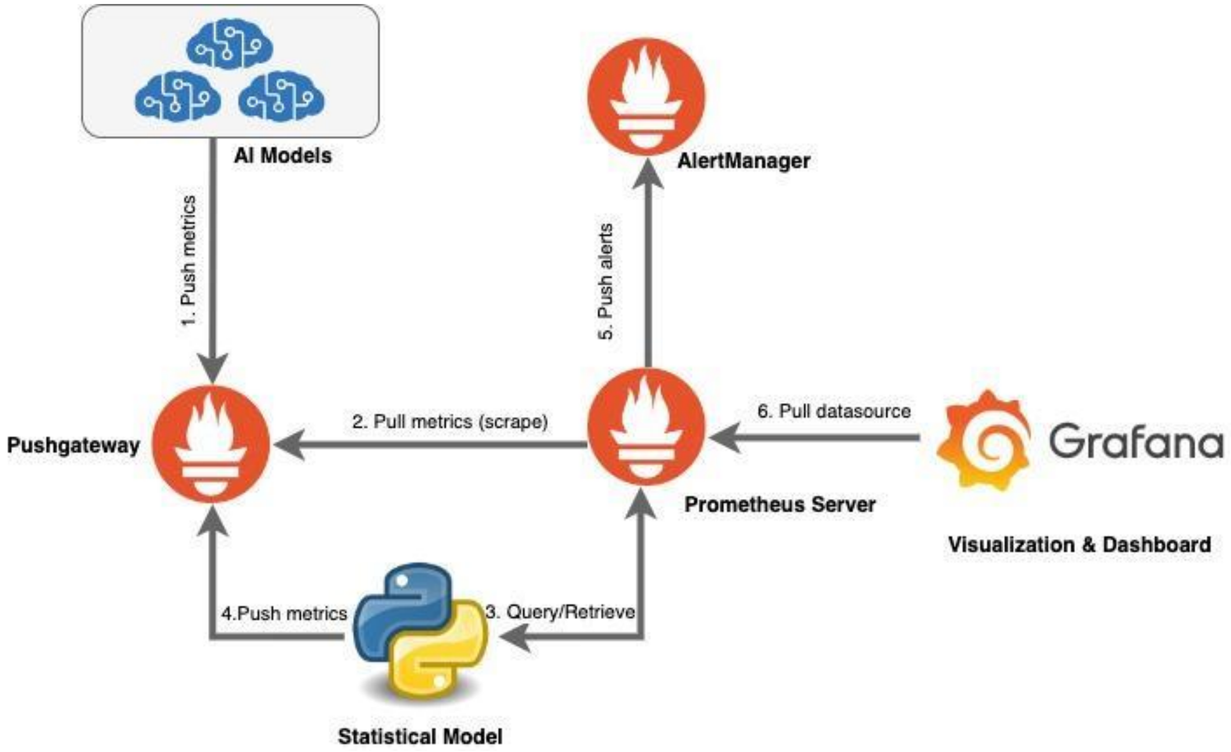


Figure 53. Pipeline for AI models metrics monitoring and visualization for drift detection in production environments

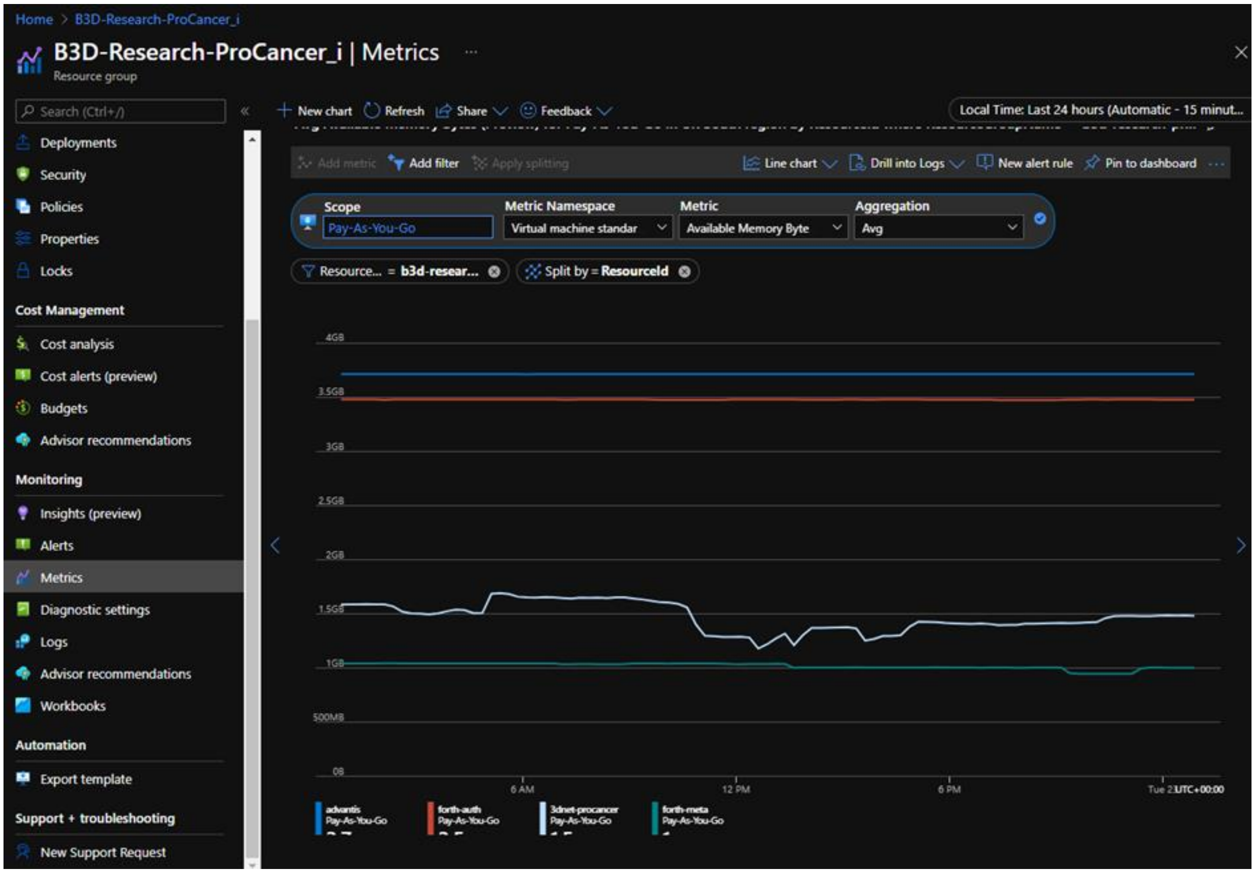
There exist different approaches to detect models drift on production. In this case, we have to consider that we do not have any ground-truth, therefore, the prediction itself needs to be monitored. Within the methods to detect model drift we can find two main groups:

[33] https://prometheus.io/

- Statistical-based approaches: statistical metrics are used to analyze differences with previous registries. This approach has been widely used in other fields, especially in finance and banking. The main statistical models used for this purpose are:
    - Population Stability Index: is a measure of population stability between two population samples.
    - Kullback-Leibler (KL) divergence: measures the difference between two probability distributions.
    - Jensen-Shannon (JS) divergence: measures the similarity between two probability distributions. Introduces some differences to KL divergence.
    - Kolmogorov-Smirnov test (KS test): non-parametric test of the equality of one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution or to compare two samples.
- Model-based approaches: a model is built to analyze the similarities between a given point or group of points and a reference.

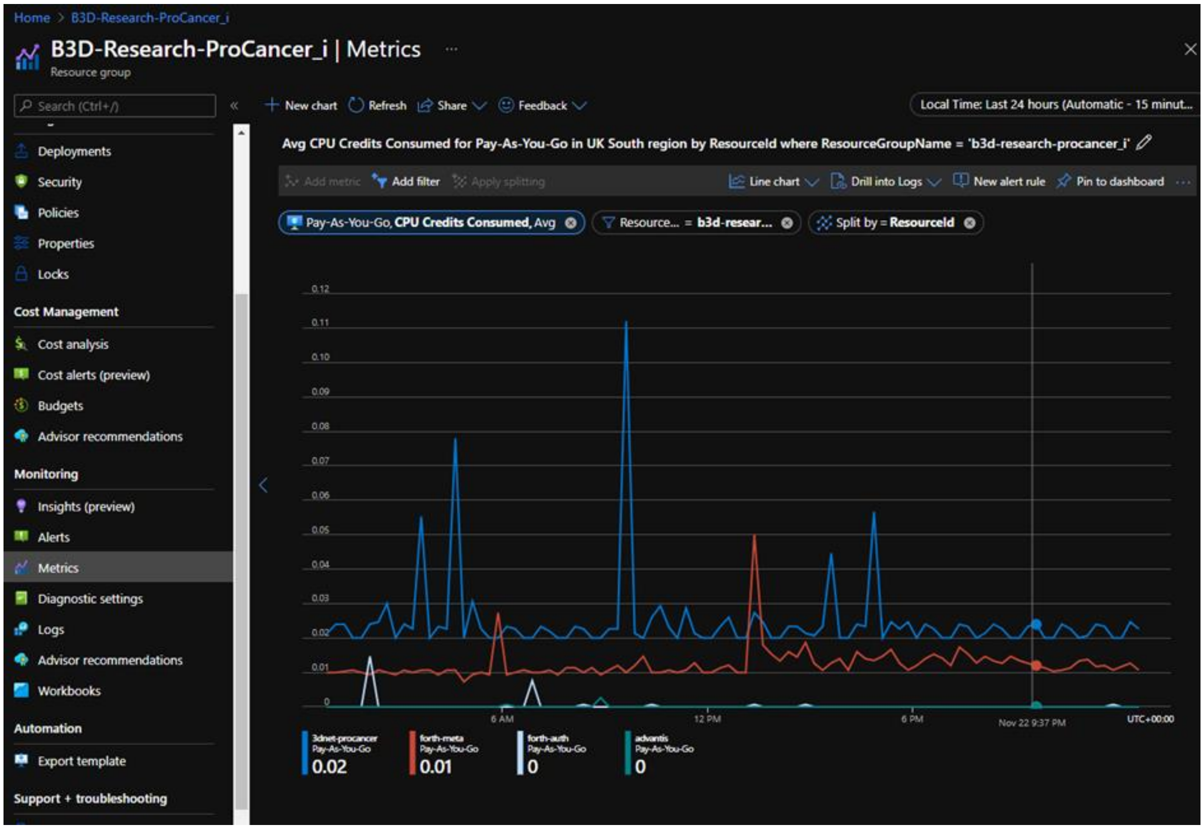
7. Platform Health Monitoring & Logging

Platform Health Monitoring & Logging is performed through the built-in monitoring functions provided by the Azure platform. This includes metrics reports about different software and virtualized hardware components as well as recommendations for improvement based on use, health and deployment characteristics of the resources allocated for the project. Logs and alerts are collected in the event of a critical error or failure that impacts the health of the physical and virtual hardware of the Virtual Machines as well as the Operating System images installed on them.



Scope: Pay-As-You-Go
Metric Namespace: Virtual machine standar
Metric: Available Memory Byte
Aggregation: Avg
Resource... = b3d-resear...
Split by = ResourceId

Figure 54. Average available memory for each VM

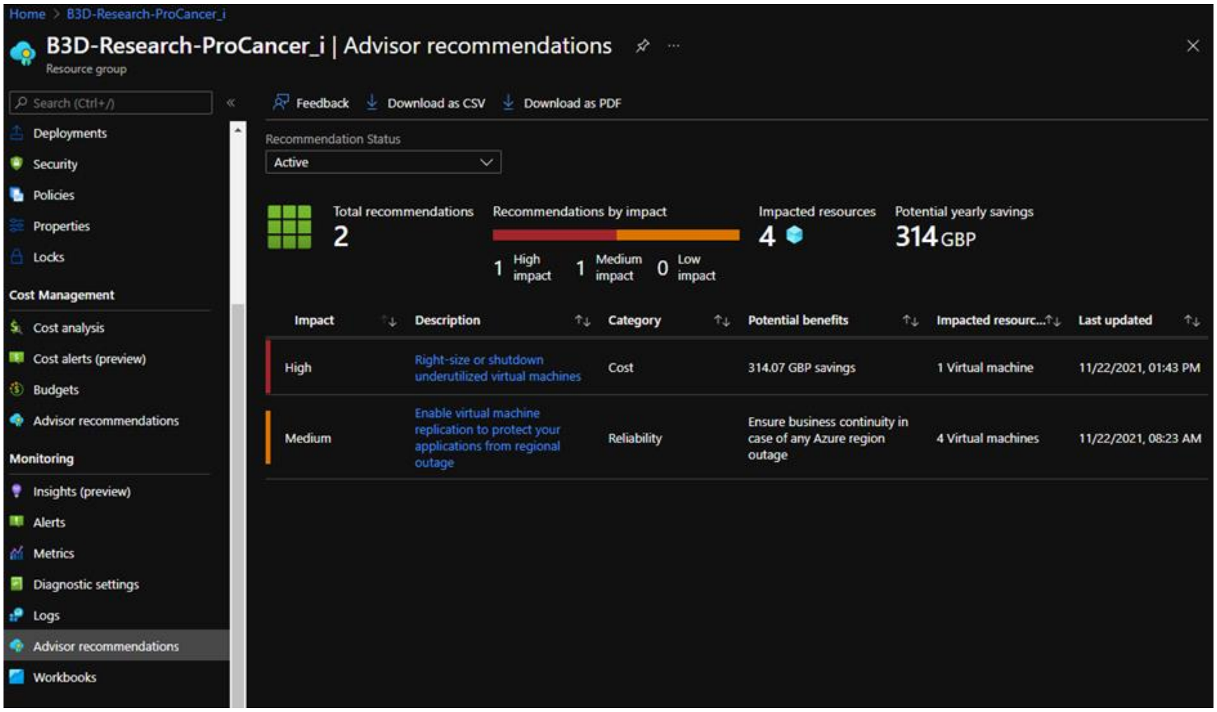


Avg CPU Credits Consumed for Pay-As-You-Go in UK South region by ResourceId where ResourceGroupName = 'b3d-research-procancer_i'

Pay-As-You-Go, **CPU Credits Consumed**, Avg | Resource... = **b3d-resear...** | Split by = **ResourceId**

3dnet-procancer Pay-As-You-Go	forth-meta Pay-As-You-Go	forth-auth Pay-As-You-Go	advantis Pay-As-You-Go
0.02	0.01	0	0

Figure 55. CPU consumed by VM



Recommendation Status: Active

Total recommendations 2 | Recommendations by impact: 1 High impact, 1 Medium impact, 0 Low impact | Impacted resources 4 | Potential yearly savings 314 GBP

Impact	Description	Category	Potential benefits	Impacted resourc...	Last updated
High	Right-size or shutdown underutilized virtual machines	Cost	314.07 GBP savings	1 Virtual machine	11/22/2021, 01:43 PM
Medium	Enable virtual machine replication to protect your applications from regional outage	Reliability	Ensure business continuity in case of any Azure region outage	4 Virtual machines	11/22/2021, 08:23 AM

Figure 56. Recommendations to improve performance and reduce cost

8. ProCancer-I Platform - UI Design Guidelines

ProstateNet, the ProCAncer-I platform, is a cloud-based platform that except for being a prostate cancer imaging archive, integrates a variety of tools, services, and modules in order to provide MLops capabilities, tools for pre-processing imaging data, an “AI model marketplace” where the developed AI Models for the different identified clinical scenarios will reside. One of the most important things in designing a UI is **consistency.**

In ProCAncer-I, in order to achieve it, the technical partners decided to define a set of rules to design the user interface of the proposed solution. In general, the UI design guidelines provide predefined templates on various design elements like:

- **Color Palette**
  List of colors that can be used
- **Typography, Headings & Labels**
  Rules for what typeface to use, font size, color, headings etc.
- **Spacing/White space**
  The rules for spacing between an element in the design.
- **Buttons**
  List of button variation, like a small button, medium button, outline button, etc.
- **Grid size & Layout Templates**
  Rules for grid size based on device and rules for ensuring layout consistency among applications (e.g. common functionalities should be found at the same place in different apps, like search, copy, paste, etc.)
- **Visuals and Media**
  How image design should be
- **Form elements**
  Like Input fields, checkboxes, dropdowns, calendar.

Initially, the team identified the Design Guidelines by the major Software Companies in the IT sector reflect the same principles regarding the design process on providing a clear overview on how to create a great User Experience focusing on **usability, utility, and desirability**.

- Fluent Design System by Microsoft at Microsoft Design[34]
- Material Design System by Google[35]
- Human Interface Guidelines by Apple[36]

Since most of the ProCAncer-I's integrated applications with a user interface will be web-based applications, the technical partners decided to start on common ground and build on defining the design guidelines to adopt the usage of a **common CSS Framework**.

[34] https://www.microsoft.com/design/fluent/

[35] https://material.io/design

[36] https://developer.apple.com/design/human-interface-guidelines/

A CSS framework is a library allowing for easier, more standards-compliant web design using the Cascading Style Sheets language. These frameworks contain a layout grid, themed UI controls, Typography templates and additional JavaScript-based functions, but are mostly design-oriented and focused on interactive UI patterns. The most well-known CSS Frameworks are:

Bootstrap
https://getbootstrap.com/
Foundation
https://get.foundation/
Bulma
https://bulma.io/
Tailwind CSS
https://tailwindcss.com/

UIkit
https://getuikit.com/
Milligram
https://milligram.io/
Pure.css (Yahoo)
https://purecss.io/
Tachyons
http://tachyons.io/

Materialize CSS
https://materializecss.com/
Skeleton CSS
http://getskeleton.com/
Semantic UI
https://semantic-ui.com/
Primer CSS (GitHub)
https://primer.style/

The partners decided to use Bootstrap as the platform's common CSS framework. The color palette was extracted by the main tools that will be integrated into the platform and are presented in the following figure (Proposed palette for UIs related to medical imaging tools - *Black Background*).

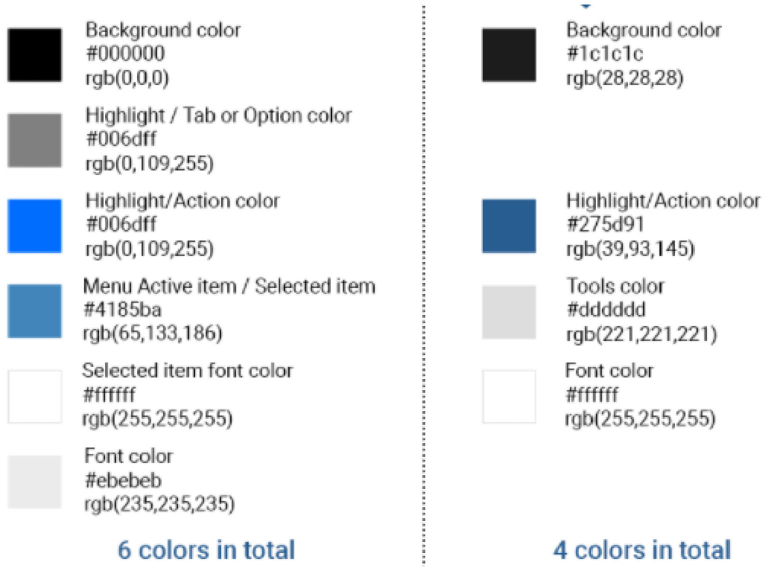


Figure 57. ProstateNet color palette

To assess the common identity and UI technical partners will define a group of front-end inspectors who will overview and manage the common identity implementation of the tools and a small group of usability inspection and accessibility testing among the clinical partners that will provide valuable hands-on insights.

9. Conclusions

This document has provided an in-depth view of the initial version of the ProCAncer-I Platform. Its scalable, fault-tolerant, and high-performance computing infrastructure has been briefly described and the various services and tools that comprise it have been detailed. Firstly, the various storage implementations were described in detail, including the DICOM Store, which is compatible with the DICOM and DICOMweb standards, the Clinical Data Repository, the Meta-Data Catalogue, which is built on top of the highly customizable Molgenis application, and the Staging Area. Afterwards, the implementation details of all platform services, both internal and user-facing, were included. The integration of the platform's Identity Provider linked to the AAI of the ELIXIR federation has been described, which is an important step towards securely enabling the retrospective upload of data by the clinical partners. Additionally, a detailed walkthrough of the workflow for imaging data anonymization and upload via the ProCAncer-I eCRF - Data Upload Tool has been provided. Moreover, the data sharing, curation, and annotation tools have been described in detail. The data sharing tool is going to enable sharing of data in a community-driven fashion on the platform, while the curation and annotation tools are going to provide the means to perform motion-correction and image co-registration and segment regions of interest, respectively. Also, the initial design and ecosystem of services/tools comprising the AI framework, which is going to allow for iterative model development, production monitoring to assess model performance, monitoring of the training and evaluation phases, as well as hardware monitoring, has been outlined. Finally, a set of common UI design guidelines has been defined in order to achieve a consistent look across the entire ProCAncer-I Platform. Based on all of the above, the alpha version of the ProCAncer-I Platform is ready to securely accept the upload of retrospective anonymized data, pending the completion of integration of more functionalities.

References

For the current deliverable, the references are denoted as footnotes.

Appendix A

Table 23. Document mapping raw ProCAncer-I data to the OMOP-CDM v6.0

	Field	Type	Concept	Mapped to table
Clinical BaselineData	patientID	String	Map to the person_source_value in the Person table.	Person
	age	int	Map to year_of_birth of the Person table. (Calculated as baseline date - age)	Person
	DRE		Digital examination of rectum - 4254766 (variable) as target_concept_id	Procedure occurrence
		negative	On rectal examination of prostate no abnormality detected - 40483554	Condition_occurrence
		positive	On rectal examination of prostate abnormality detected - 43531580	Condition_occurrence
	totalPSAvalue	real	Total PSA level - 44793131, 8842 - nanogram per milliliter as unit_concept_id	Measurement
	freePSAvalue	real	Free prostate specific antigen level - 4194418, 8842 - nanogram per milliliter as unit_concept_id	Measurement
	ratioPSAvalue	real	Free:total PSA ratio 4215704, % - 8554 as unit_concept_id	Measurement
	baseline	bool	if true add: Baseline Visit - 2000000007, if false add: PSA visit - 2000000008	Visit_occurrence
	prostateVolume	real	Prostate Volume - 2000000022, mm3 - 8587 as unit_concept_id.	Measurement
	useCaseType	String	Map to Note table	Note
Family History	relativePCA	int	if "yes": Family history of clinical finding - 4167217, "no": No family history of clinical finding- 4051104, "unknown": Family history unknown:4236282 , value_as_concept_id: 4163261 Malignant tumor of prostate	Observation

	Field	Type	Concept	Mapped to table
	relativeBCa	int	if "yes": Family history of clinical finding - 4167217, "no": No family history of clinical finding- 4051104, "unknown": Family history unknown:4236282 as target_concept_id and value_as_concept_id: 4112853 Malignant tumour of breast	Observation
	relativeAggressiveCancer	int	if "yes": Family history of clinical finding - 4167217, "no": No family history of clinical finding- 4051104, "unknown": Family history unknown:4236282 as target_concept_id and value_as_concept_id: 2000000006 - Aggressive Cancer	Observation
	relativePCAage	int	Mapped to Note table. note_title as "age", note_text as the age value.	Note
	relativeBCaAge	int	Mapped to Note table. note_title as "age", note_text as the age value.	Note
	relativeAggressiveCancer	int	Mapped to Note table. note_title as "age", note_text as the age value.	Note
Imaging Info	daysAfterBaselineMRI	int	Mapped to procedure_date for MRI	Procedure occurrence
	MRI positive	boolean	Multiparametric MRI of prostate - 36714087	Procedure occurrence
			MRI scan abnormal - 4059669 if true, or MRI scan normal - 4058339 if false	Condition occurrence
Biopsy Info	biopsyperformed	bool	Biopsy of prostate - 4278515	Procedure occurrence
	biopsyPositive	bool	Biopsy result abnormal - 4013824 if positive, or Biopsy result normal - 4012811 if negative	Condition occurrence
	biopsyType	enum	Ultrasound guided biopsy - 4082629, MRI-US fusion guided prostate biopsy - 37206816, Inboard biopsy - 2000000005	Procedure occurrence
Lesion Info	diam1	int	Lesion size, greatest dimension - 4245121, millimeter - 8588 as unit_concept_id	Measurement

	Field	Type	Concept	Mapped to table
	diam2	int	Lesion size, additional dimension - 4245121, millimeter - 8588 as unit_concept_id	Measurement
	volume	int	Tumor volume - 4121185, mm3 - 8686 as unit_concept_id.	Measurement

	Field	Type	Concept	Mapped to table
	location	int	Site of lesion - 4135405 (variable) as target_concept_id, following as value: "BLAS": 37396905, "BRAS": 37396904, "BLTZa": 37396909, "BRTZa": 37396908, "BLTZp": 37396910, "BRTZp": 37399562, "BLCZ": 37396912, "BRCZ": 37396911, "BLPZa": 37396915, "BRPZa": 37399563, "BLPZpl": 37396917, "BRPZpl": 37396916, "MLAS": 37396920, "MRAS": 37396919, "MLTZa": 37396923, "MRTZa": 37396922, "MLTZp": 37396925, "MRTZp": 37396924, "MLPZa": 37396928, "MRPZa": 37396927, "MLPZpl": 37396930, "MRPZpl": 37396929, "MLPZpm": 37396932, "MRPZpm": 37396931, "ALAS": 37396934, "ARAS": 37396933, "ALTZa": 37396938, "ARTZa": 37396937, "ALTZp": 37396940, "ARTZp": 37396939, "ALPZa": 37399019, "ARPZa": 37399567, "ALPZpl": 37396944, "ARPZpl": 37396943, "ALPZpm": 37396945, "ARPZpm": 37399568	Measurement

	Field	Type	Concept	Mapped to table
	X, Y, Z	int	overlapping lesion - 36770647, x- 4223576, y- 4227659, z - 4221919	Measurement
	PI-RADS	int	There is no standard OMOP concept for PI-RADS. Map to custom concept "PI-RADS score" - 2000000004	Measurement
	gleason1	int	Primary Gleason pattern - 4293445	Measurement
	gleason2	int	Secondary Gleason pattern - 4297948	Measurement
			Gleason's Score on Prostatectomy/Autopsy - 35918021 (variable) as target_concept_id (there are many concepts as values for all the combinations of the primary and secondary)	Measurement
	MRIVisible		"MRI_visible"- 2000000002 connected to the lesion condition	Observation
	indexLesion		"index lesion" - 2000000001 connected to the lesion condition	Observation
	prostatectomyBasedLocation	boolean	Note_title: lesionLocationBasedOn, note_text: either Biopsy or Prostatectomy	Note
Prostatectomy Info	prostatectomy_performed	int	Prostatectomy - 4235738	Procedure occurrence
	RPmethod	enum	Radical Retropubic prostatectomy - 4276520, Robot assisted laparoscopic radical prostatectomy - 46270921, Laparoscopic prostatectomy - 44809585, , Radical Perineal prostatectomy - 4338373, Nerve sparing is missing - 2000000003, Other: 2003983	Procedure occurrence

	Field	Type	Concept	Mapped to table
Stage-TNM	T	String	Prostate Cancer cTX TNM Finding by AJCC/UICC 8th edition: 1538985, Prostate Cancer cT0 TNM Finding by AJCC/UICC 8th edition: 1538044, Prostate Cancer cT1 TNM Finding by AJCC/UICC 8th edition: 1538812, Prostate Cancer cT1a TNM Finding by AJCC/UICC 8th edition: 1538366, Prostate Cancer cT1b TNM Finding by AJCC/UICC 8th edition: 1538219, Prostate Cancer cT1c TNM Finding by AJCC/UICC 8th edition: 1538775, Prostate Cancer cT2 TNM Finding by AJCC/UICC 8th edition:1538375 , Prostate Cancer cT2a TNM Finding by AJCC/UICC 8th edition: 1539231, Prostate Cancer cT2b TNM Finding by AJCC/UICC 8th edition: 1538566, Prostate Cancer cT2c TNM Finding by AJCC/UICC 8th edition: 1538629, Prostate Cancer cT3 TNM Finding by AJCC/UICC 8th edition: 1538400, Prostate Cancer cT3a TNM Finding by AJCC/UICC 8th edition: 1539253, Prostate Cancer cT3b TNM Finding by AJCC/UICC 8th edition: 1539229, Prostate Cancer cT4 TNM Finding by AJCC/UICC 8th edition: 1538855	Measurement
	N	String	Prostate Cancer pNX TNM Finding by AJCC/UICC 8th edition: 1537945, Prostate Cancer pN0 TNM Finding by AJCC/UICC 8th edition: 1538383, Prostate Cancer pN1 TNM Finding by AJCC/UICC 8th edition: 1538914	Measurement
	M	String	Prostate Cancer cM0 TNM Finding by AJCC/UICC 8th edition: 1538822, Prostate Cancer cM1 TNM Finding by AJCC/UICC 8th edition: 1538290, Prostate Cancer cM1a TNM Finding by AJCC/UICC 8th edition: 1539119, Prostate Cancer cM1b TNM Finding by AJCC/UICC 8th edition: 1538995, Prostate Cancer cM1c TNM Finding by AJCC/UICC 8th edition: 1538321, TNM Prostate tumor staging - 4110284 (variable) as target_concept_id, following as value: MX category - 4098852	Measurement

	Field	Type	Concept	Mapped to table
Cancer Diagnosis	carcinomaClass	String	Map from text to ICD03 then to OMOP,"Acinar adenocarcinoma" : "8140/3-C61.9", "Intraductal carcinoma" : "8500/2-C61.9", "Ductal adenocarcinoma" : "8500/3-C61.9", "Urothelial carcinoma" : "8120/3-C61.9", "Adenosquamous carcinoma" : "8560/3-C61.9", "Squamous cell carcinoma" : "8070/3-C61.9", "Basal cell carcinoma" : "8147/3-C61.9", "Adenocarcinoma with neuroendocrine differentiation": "8574/3-C61.9", "Small cell neuroendocrine carcinoma": "8041/3-C61.9", "Large cell neuroendocrine carcinoma": "8013/3-C61.9", "8140/3-C61.9": 44499685,"8500/2-C61.9": 36556105, "8500/3-C61.9": 44499514, "8120/3-C61.9": 44501195, "8560/3-C61.9": 44503558, "8070/3-C61.9": 44499428, "8147/3-C61.9": 36554023, "8041/3-C61.9": 44502766, "8013/3-C61.9": 44499864, 44499685: 4161028, 36556105: 36556105, 44499514: 44499514, 44501195: 44501195, 44503558: 44503558, 44499428: 4164017, 36554023: 36554023, 44502766: 44502766, 44499864: 44499864	Condition occurrence
Invasion Info	resectionMarginsStatus	enum	Resection Margin Involved by tumor - 36768316, Resection Margin Uninvolved by tumor - 36770153	Measurement - Cancer Modifier
	extraProstaticExtension	bool	Extraprostatic Extension (EPE): 6768810, Present: 4181412, Absent: 4132135	Measurement - Cancer Modifier
	wheelerRanking		Prostatic Capsular Invasion (PCI) level by Wheeler ranking - 2000000021	Measurement - Cancer Modifier
	perineuralInvasion	bool	Perineural Invasion: 36768846, Present: 4181412, Absent: 4132135	Measurement - Cancer Modifier

	Field	Type	Concept	Mapped to table
	seminalVesicalInvasion	bool	Invasion into the Seminal vesicle: 36769282, Present: 4181412, Absent: 4132135	Measurement - Cancer Modifier
Treatment Regimen	Radiotherapy technique	enum	IMRT - intensity modulated radiation therapy - 40480519, IGRT (image-guided radiation therapy) - 45956251, 3D CRT (three dimensional conformal radiation therapy) - 45911882, HSR: (hypofractionated stereotactic RT) - 44809885 Other: 3169882	Treatment Regimen, Episode, Procedure
	Hormonotherapy	bool	Hormonotherapy - 35803407	Treatment Regimen, Episode, Procedure
	Total dose	int	Total dose - 35918606, unit_concept_id: Gray - 9519	Measurement
	fractionsWeek	int	Phase I Number of Fractions: 35918481, unit_concept_id: 44777559- perWeek	Measurement
	doseFraction	int	Phase I Dose per Fraction: 35918531, unit_concept_id: Gray - 9519	Measurement
	rectalToxAcute	int	RTOG/EORTC rectal acute radiation morbidity score: 2000000023	Measurement
	rectalToxChronic	int	RTOG/EORTC rectal chronic radiation morbidity score: 2000000024	Measurement
	genitourinaryToxAcute	int	RTOG/EORTC genitourinary acute radiation morbidity score: 2000000025	Measurement
	genitourinaryToxChronic	int	RTOG/EORTC genitourinary acute radiation morbidity score: 2000000026	Measurement
Follow-Up	fupPSANormalRange	bool	PSA normal: 4013978, PSA abnormal: 439453	Condition_occurrence
	fupMRINegative	bool	MRI normal: 4058339, abnormal: 4059669	Condition_occurrence
	fupBiopsyNegative	bool	Biopsy normal: 4012811, abnormal: 4013824	Condition_occurrence
	monthsFinalDiagnosis	int	condition_occurrence_datetime	Condition_occurrence

	Field	Type	Concept	Mapped to table
	diagnosisTNM		check TNM above	Measurement
	imagingPerformed		check imaging above	Procedure_occurrence
	imagingType		check imaging above	Procedure_occurrence
	imgMetastasis description		Map to measurement_source_value with measurement concept id depending on the metastasis site	Measurement
	totalPSAvalue	real	Total PSA level - 44793131, 8842 - nanogram per milliliter as unit_concept_id	Measurement
	freePSAvalue	real	Free prostate specific antigen level - 4194418, % - 8554 as unit_concept_id	Measurement
	ratioPSAvalue	real	Free:total PSA ratio 4215704	Measurement
	psadate	datetime	Map to measurement_date	Measurement
QoL questionnaire	epic26	int	Expanded Prostate cancer Index Composite score (EPIC-26): 2000000033	Measurement
	EORTC_Q52	int	2000000034 - Q52. To what extent was sex enjoyable for you?	Observation
	EORTC_Q53	int	2000000035 - Q53. Did you have difficulty getting or maintaining an erection?	Observation
	EORTC_Q54	int	2000000036 - Q54. Did you have ejaculation problems (e.g. dry ejaculation)?	Observation
	EORTC_Q55	int	2000000037 - Q55. Have you felt uncomfortable about being sexually intimate?	Observation

Table 24. ProCAncer-I custom vocabulary

concept_id	concept_name	domain_id	vocabulary_id	concept_class_id	standard_concept	concept_code	valid_start_date	valid_end_date
2000000001	Index Lesion	Observation	PROCANCERI	Study Variable	S	1	1970-01-01	2099-12-31
2000000002	Lesion Visible on MRI	Observation	PROCANCERI	Study Variable	S	2	1970-01-01	2099-12-31
2000000003	Nerve-sparing prostatectomy	Procedure	PROCANCERI	Study Variable	S	3	1970-01-01	2099-12-31
2000000004	PI-RADS score	Measurement	PROCANCERI	Oncology Variable	S	4	1970-01-02	2100-01-01
2000000005	In-Bore Biopsy	Procedure	PROCANCERI	Study Variable	S	5	1970-01-01	2099-12-31
2000000006	Aggressive Cancer	Condition	PROCANCERI	Study Variable	S	6	1970-01-01	2099-12-31
2000000007	Baseline Visit	Visit	PROCANCERI	Visit	S	7	1970-01-01	2099-12-31
2000000008	PSA Visit	Visit	PROCANCERI	Visit	S	8	1970-01-01	2099-12-31
2000000009	Baseline Visit - Imaging	Visit	PROCANCERI	Visit	S	9	1970-01-01	2099-12-31
2000000010	Follow-Up Visit	Visit	PROCANCERI	Visit	S	10	1970-01-01	2099-12-31
2000000011	Follow-Up Visit - Baseline	Visit	PROCANCERI	Visit	S	11	1970-01-01	2099-12-31
2000000012	Follow-Up Visit - PSA	Visit	PROCANCERI	Visit	S	12	1970-01-01	2099-12-31
2000000013	Follow-Up Visit - Biopsy	Visit	PROCANCERI	Visit	S	13	1970-01-01	2099-12-31
2000000014	Follow-Up Visit - Imaging	Visit	PROCANCERI	Visit	S	14	1970-01-01	2099-12-31
2000000015	L0	Measurement	PROCANCERI	Oncology Variable	S	15	1970-01-01	2099-12-31
2000000016	L1	Measurement	PROCANCERI	Oncology Variable	S	16	1970-01-01	2099-12-31

2000000017	L2	Measurement	PROCANCERI	Oncology Variable	S	17	1970-01-01	2099-12-31
2000000018	L3F	Measurement	PROCANCERI	Oncology Variable	S	18	1970-01-01	2099-12-31
2000000019	L3E	Measurement	PROCANCERI	Oncology Variable	S	19	1970-01-01	2099-12-31
2000000020	LX	Measurement	PROCANCERI	Oncology Variable	S	20	1970-01-01	2099-12-31
2000000021	Wheeler Ranking	Measurement	PROCANCERI	Oncology Variable	S	21	1970-01-01	2099-12-31
2000000022	Prostate Volume	Measurement	PROCANCERI	Study Variable	S	22	1970-01-01	2099-12-31
2000000023	RTOG/EORTC rectal acute radiation morbidity score	Measurement	PROCANCERI	Oncology Variable	S	23	1970-01-02	2100-01-01
2000000024	RTOG/EORTC rectal chronic radiation morbidity score	Measurement	PROCANCERI	Oncology Variable	S	24	1970-01-02	2100-01-01
2000000025	RTOG/EORTC genitourinary acute radiation morbidity score	Measurement	PROCANCERI	Oncology Variable	S	25	1970-01-02	2100-01-01
2000000026	RTOG/EORTC genitourinary chronic radiation morbidity score	Measurement	PROCANCERI	Oncology Variable	S	26	1970-01-02	2100-01-01
2000000027	Grade 1	Meas Value	PROCANCERI	Oncology Variable	S	27	1970-01-02	2100-01-01
2000000028	Grade 2	Meas Value	PROCANCERI	Oncology Variable	S	28	1970-01-03	2100-01-02
2000000029	Grade 3	Meas Value	PROCANCERI	Oncology Variable	S	29	1970-01-04	2100-01-03

2000000030	Grade 4	Meas Value	PROCANCERI	Oncology Variable	S	30	1970-01-04	2100-01-03
2000000031	Grade 5	Meas Value	PROCANCERI	Oncology Variable	S	31	1970-01-04	2100-01-03
2000000032	Grade 5	Meas Value	PROCANCERI	Oncology Variable	S	32	1970-01-04	2100-01-03
2000000033	Expanded Prostate cancer Index Composite score (EPIC-26)	Measurement	PROCANCERI	Study Variable	S	33	1970-01-05	2100-01-04
2000000034	Q52. To what extent was sex enjoyable for you?	Observation	PROCANCERI	Study Variable	S	34	1970-01-06	2100-01-05
2000000035	Q53. Did you have difficulty getting or maintaining an erection?	Observation	PROCANCERI	Study Variable	S	35	1970-01-07	2100-01-06
2000000036	Q54. Did you have ejaculation problems (e.g. dry ejaculation)?	Observation	PROCANCERI	Study Variable	S	36	1970-01-08	2100-01-07
2000000037	Q55. Have you felt uncomfortable about being sexually intimate?	Observation	PROCANCERI	Study Variable	S	37	1970-01-09	2100-01-08
2000000038	EORTC Quality of Life Questionnaire-PR25	Measurement	PROCANCERI	Study Variable	S	38	1970-01-10	2100-01-09

Appendix B

Here we show samples of the metadata generated by the Curation Tool for each curation function and stored on the Metadata Catalogue. At a high level, each sample includes (a) the name of the curation function: *motion_correction* or *coregistration*, (b) the UIDs of the series involved, and (c) the hyperparameters used to configure the underlying registration algorithm.

```
{
  "curation_function": "motion_correction",
  "meta": {
    "source_series_uid": "1.3.6.1.4.1.19291.2.1.2.12482551947010615361365978484506"
  },
  "parameters": {
    "registration_steps": [
      {
        "levels": [
          {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "translation"
      },
      {
        "levels": [
          {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "rigid"
      },
      {
        "levels": [
          {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "affine"
      }
    ],
    "nbins": 32,
    "sampling_prop": 0.1
  }
}
```

```
{
  "curation_function": "coregistration",
  "meta": {
    "source_series_uid": "1.3.6.1.4.1.19291.2.1.2.839069178662379942206215885338",
    "static_series_uid": "1.3.6.1.4.1.19291.2.1.2.1248255361365979160792"
  },
  "parameters": {
    "registration_steps": [
      {
        "levels": [
          {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "translation"
      },
      {
        "levels": [
         {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "rigid"
      },
      {
        "levels": [
          {
            "factor": 4,
            "level_iters": 500,
            "sigma": 3
          },
          {
            "factor": 2,
            "level_iters": 100,
            "sigma": 2
          },
          {
            "factor": 1,
            "level_iters": 10,
            "sigma": 0
          }
        ],
        "registration_step_name": "affine"
      }
    ],
    "nbins": 32,
    "sampling_prop": 0.1
  }
}
```

Appendix C

Testing the API

The ProCAncer-I Image Repository API is provided at:

https://procancer-i.3dnetmedical.com:63801

The API can be tested using applications such as cURL, Postman or via standard RESTful calls from multiple programming languages.

cURL tutorial:

https://github.com/microsoft/dicom-server/blob/main/docs/tutorials/use-dicom-web-standard-apis-with-curl.md

Python tutorial:

https://github.com/microsoft/dicom-server/blob/main/docs/tutorials/use-dicom-web-standard-apis-with-python.md

Python code with `StudyID` already in the repository:

```
import requests
from pathlib import Path
from urllib3.filepost import encode_multipart_formdata, choose_boundary

base_url = "https://procancer-i.3dnetmedical.com:63801"
study_uid = "1.3.6.1.4.1.14519.5.2.1.7311.5101.158323547117540061132729905711"

client = requests.session()

url = f'{base_url}/studies'
headers = {'Accept':'application/dicom+json'}
params = {'StudyInstanceUID':study_uid}

response = client.get(url, headers=headers, params=params) #, verify=False)

print(response.json())
```