Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220446

Data Ingestion for AI in Prostate Cancer

Haridimos KONDYLAKIS[a], Stelios SFAKIANAKIS[a], Varvara KALOKYRI[a], Nikolaos TACHOS[a], Dimitrios FOTIADIS[a], Kostas MARIAS[a] and Manolis TSIKNAKIS[a]
[a] *FORTH-ICS, N Plastira 100, Heraklion, Crete, Greece*

Abstract. Prostate cancer (PCa) is one of the most prevalent cancers in the male population. Current clinical practices lead to overdiagnosis and overtreatment necessitating more effective tools for improving diagnosis, thus the quality of life of patients. Recent advances in infrastructure, computing power and artificial intelligence enable the collection of tremendous amounts of clinical and imaging data that could assist towards this end. ProCAncer-I project aims to develop an AI platform integrating imaging data and models and hosting the largest collection of PCa (mp)MRI, anonymized image data worldwide. In this paper, we present an overview of the overall architecture focusing on the data ingestion part of the platform. We describe the workflow followed for uploading the data and the main repositories for storing imaging data, clinical data and their corresponding metadata.

Keywords. Data Ingestion, Prostate Cancer

1. Introduction

About 1,300,000 citizens of the European Union are estimated to have had a prostate cancer diagnosis in the last five years [1]. The severe socioeconomic burden for health services and the negative effects on the quality of life of patients call for immediate actions [2][3]. Artificial Intelligence on the other hand has the potential to bring medicine from the era of 'sick care' to the era of healthcare and prevention, fueled by the availability of large datasets ("big data"), substantial advances in computing power, and new deep-learning algorithms [4]. However, the availability of large, quality-controlled datasets for building those AI models, currently remains a major challenge. To this purpose, several health imaging repositories have been created [5][6], such as the Cancer Imaging Archive (TCIA) [7]. However, the vast majority of these repositories have been created as stand-alone entities, being currently not in a position to become interoperable with similar existing initiatives. The need for the creation of a fully FAIR (Findable, Accessible, Interoperable, Reusable), GDPR compliant, European imaging repository still stands [8] and has been recognized by other EU research projects like PRIMAGE [9] and CHAIMELEON [10] however, still there are not tangible results from these projects. ProCAncer-I's vision is to become a catalyst in this process by creating the first European, ethical- and GDPR (General Data Protection Regulation) compliant, quality-controlled, prostate cancer (PCa) related, medical imaging platform, in which both large-scale data and AI algorithms will co-exist. To this end, the ProstateNet dataset featuring an unprecedented 1.5 million image representations of prostate cancer will be created within a sustainable AI cloud-based platform for the development, implementation, verification and validation of trustworthy, usable and reliable AI models. In this paper

we provide a glimpse of the overall architecture focusing on the data ingestion part of the platform. model used.

2. Architecture

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. The logical view of the ProCAncer-I platform with the main domain specific areas of functionality of the system is shown in Figure 1. 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".

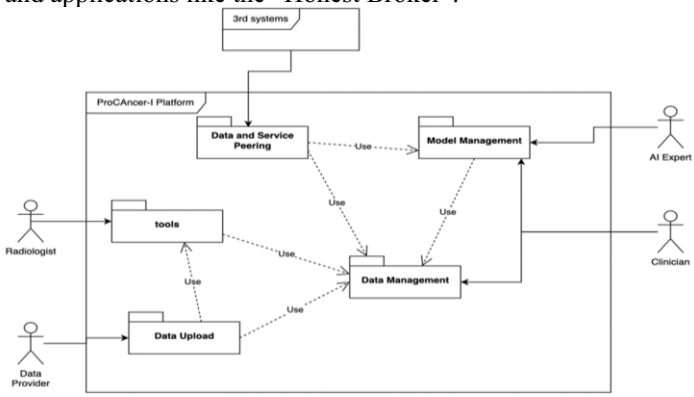

Figure 1. The main subsystems of the ProCAncer-I platform

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. In the sequel we present data ingestion and upload, data management and domain specific tools that have been developed and integrated to allow data providers to make their data sets available to the ProCAncer-I community securely and fully annotated.

Data ingestion and upload. 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. The ProCAncer-I cloud platform storage, ProstateNet, is comprised of 3 components: a) the DICOM Object Store which stores medical imaging data; b) the clinical data document store which stores the clinical data; and c) the meta-data catalogue which stores metadata and semantic annotations to enable rich search and discovery of data and its exploitation.

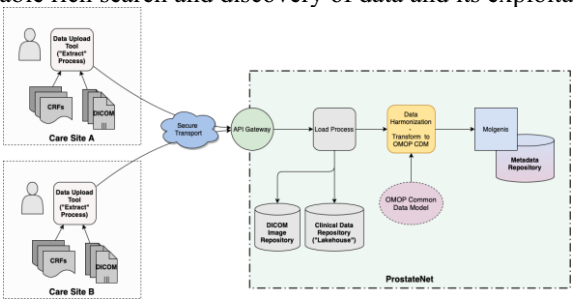

Figure 2. Main components and processes of the data ingestion pipeline

The whole data ingestion pipeline with its main components and subprocesses is shown in Figure 2. The clinical partners 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 is 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").

DICOM Image Repository. The ProCAncer-I project DICOM Image Repository provides the necessary services for saving, updating, and retrieving DICOM studies. The implementation of the repository is compliant with both the DICOM and HL7 standards, thus allowing seamless interoperability with existing PACS systems and scanners. To support the several steps of data curation, annotation and AI research and development, the repository provides mechanisms for querying and retrieving data through an API gateway using the standard DICOMWeb [11] suite of programming interfaces.

Clinical Data Repository. The Clinical Data Repository is the central data warehouse of the ProCAncer-I platform, accessible through a web-based ("RESTful") API. In the data upload process this repository stores the submitted clinical and image related information before the metadata extraction and their persistence in the Metadata Catalogue. In this way, the data ingestion follows an Extract-Load-Transform (ELT) [12] design where this repository is responsible for maintaining the data in their original submitted format. The Clinical Repository is similar to a "data lake" that contains all of the uploaded information (except the imaging data) in the format that was uploaded. Being similar to a traditional data warehouse it offers transactional interactions and both structured and unstructured data storage. In addition, 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 DICOM 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 also 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. Finally, it maintains an upload log so to enable traceability, by gathering of statistics and monitoring of the use of the data platform are possible.

Meta-Data Catalogue. In ProCAncer-I, the MOLGENIS (https://www.molgenis.org/) platform has been adopted to serve as the main metadata catalogue of the project, whereas OMOP-CDM and its extensions are used as the common data model to store the available metadata. MOLGENIS has a completely customized data system allowing modeling of the data using external data models. 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.

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. It supports the standardization and harmonization of health data as well as 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, focusing on including annotation, segmentation and curation data as radiomics.

OMOP-CDM ETL. To get from the native/raw data provided by the clinical sites 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. 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. 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. In addition, we have designed an initial schema for storing the image curation information. During the whole pipeline the users are able to observe the various steps and specific tools for quality control are available. These tools include visualization interfaces for the imaging data and segmentation masks, as well as for the ETL output and the OMOP-CDM mapping and storage in MOLGENIS.

Deployment. Apart from the functional requirements and the data ingestion pipeline presented above, the envisaged platform needs to address certain non-functional requirements, such as scalability, security, availability, and overall performance. To address these needs, the platform is deployed on a commercial cloud using "cloud-native" technologies and tools. In particular, all the platform's components are deployed as containers using Kubernetes as the container orchestration platform for automated container deployment. The platform network is also "containerized" with both control and data plane composed of microservices with flexible deployment specifications to address fluctuations in workload. Finally, there's built-in observability and analytics functionality in order to enable continuous monitoring and automated troubleshooting

during the upload and data transformation phases, using existing monitoring tools like Prometheus (https://prometheus.io/) and Grafana (https://grafana.com/).

3. Conclusions

This paper provided an overview of the initial version of the ProCAncer-I Platform data ingestion modules. The various storage modules were described, including the DICOM Image Repository, which is compatible with the DICOM and DICOMweb standards, the Clinical Data Repository and the Meta-Data Catalogue, which is built on top of the highly customizable MOLGENIS application. 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. As next steps the data ingestion platform will be extensively evaluated by the users and also the ecosystem of services/tools comprising the AI framework is currently designed and under implementation. Those will be reported in a follow-up paper.

Acknowledgements

Research reported in this paper has been supported from the ProCAncer-I H2020 project (No 952159) and has received funding from the European Union's Horizon 2020 research and innovation programme.

References

[1] Epidemiology of prostate cancer in Europe: https://ec.europa.eu/jrc/en/publication/epidemiology-prostate-cancer-europe.
[2] Kondylakis H, et al. Patient empowerment for cancer patients through a novel ICT infrastructure. Journal of biomedical informatics 101 (2020): 103342.
[3] Kondylakis H, et al. Developing a data infrastructure for enabling breast cancer women to BOUNCE back. IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE, 2019.
[4] Thrall JH, et al. Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success, Journal of the American College of Radiology, 2018; 15(3):504-508.
[5] Sudlow C, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med (2015) 12:e1001779.
[6] Bamberg F, et al. Whole-Body MR Imaging in the German National Cohort: Rationale, Design, and Technical Background. Radiology (2015) 277:206–20.
[7] Clark K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7
[8] Gabelloni M, et al. Bridging Gaps Between Images and Data: A Systematic Update on Imaging Biobanks. Eur Radiol (2021). doi: 10.1007/s00330-021-08431-6
[9] Martí-Bonmatí L, et al. PRIMAGE Project: Predictive In Silico Multiscale Analytics to Support Childhood Cancer Personalised Evaluation Empowered by Imaging Biomarkers. Eur Radiol Exp (2020) 4:22.
[10] Bonmatí LM, et al. CHAIMELEON Project: Creation of a Pan-European Repository of Health Imaging Data for the Development of AI-Powered Cancer Management Tools. Frontiers in Oncology. 2022:515.
[11] Genereaux BW, et al. DICOMweb™: Background and Application of the Web Standard for Medical Imaging. J Digit Imaging. 2018 Jun;31(3):321–326.
[12] Cohen J, Dolan B, Dunlap M, Hellerstein JM, Welton C. MAD skills: new analysis practices for big data. Proc VLDB Endow. 2009 Aug;2(2):1481–1492.