Standardizing Data and Metadata: Experiences from three AI4HI Projects

Haridimos Kondylakis, Computer Science Department, University of Crete & Institute of Computer Science, FORTH, N. Plastira 100, GR70013, Heraklion, Crete, Greece, kondylak@ics.forth.gr

Varvara Kalokyri, Institute of Computer Science, FORTH, N. Plastira 100, GR70013, Heraklion, Crete, Greece, vkalokyri@ics.forth.gr

Alexandra Kosvyra, School of Medicine, Aristotle University of Thessaloniki, 54124, Greece, aekosvyra@auth.gr

Pedro Mallol, La Fe Health Research Institute, Avda. Fernando Abril Martorell, 106, 46026 Valencia, Spain, pedro_mallol@iislafe.es

Stelios Sfakianakis, Institute of Computer Science, FORTH, N. Plastira 100, GR70013, Heraklion, Crete, Greece, ssfak@ics.forth.gr

Sara Colantonio, National Research Council, Institute of Information Science and Technologies (ISTI) “Alessandro Faedo”, Via Moruzzi, 1, Pisa, 56127, Italy, sara.colantonio@isti.cnr.it

Dimitrios I. Fotiadis, Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, Ioannina, Greece fotiadis@uoi.gr

Kostas Marias, Institute of Computer Science, FORTH, N. Plastira 100, GR70013, Heraklion, Crete, Greece, kmarias@ics.forth.gr

Manolis Tsiknakis, Institute of Computer Science, FORTH, N. Plastira 100, GR70013, Heraklion, Crete, Greece, tsiknaki@ics.forth.gr

Abstract.

In the realm of modern healthcare, data science plays a pivotal role, offering a multitude of benefits to both patients and medical professionals. However, to harness the benefits of the wealth of data now available, the data should be homogenized and accessible. Data standardization ensures uniformity and consistency across different diagnostic modalities whereas metadata, provides essential context, encompassing information about data provenance, acquisition protocols, patient demographics, and study parameters. In this paper we present the standardizing approaches of three EU projects focusing on cancer,

demonstrating the decisions taken and the workflows adopted. Then we present experiences and problems offering a valuable resource for other similar approaches in the future.

Table of abbreviations

AI	Artificial Intelligence
AI4HI	Artificial Intelligence for Health Imaging
eCRF	Electronic Clinical Report Form
EU	European Union
CDM	Common Data Model
DICOM	Digital Imaging and Communications in Medicine
OMOP	Observational Medical Outcomes Partnership
PACS	Picture archiving and communication system
PCa	Prostate Cancer
RWD	Real World Data

1. Introduction

In the dynamic landscape of modern healthcare, where diagnostic techniques have evolved rapidly, the generation of copious amounts of data has become commonplace. These data span diverse domains, including clinical imaging, pathology, and genomic information. However, the true potential of this wealth of data lies in its accessibility, standardization, and effective utilization. In this context, two critical pillars emerge: data standardization and metadata.

Data standardization ensures uniformity and consistency across different diagnostic modalities. When imaging data and clinical records adhere to standardized formats, researchers can seamlessly collaborate, compare findings, and develop robust predictive models. For instance, standardized radiomic features allow for meaningful comparisons between different imaging studies, enabling precise disease characterization. Without such standardization, the reproducibility of research findings and the traceability of sample characteristics would be compromised [1][2].

Metadata, on the other hand, provides essential context. It encompasses information about data provenance, acquisition protocols, patient demographics, and study parameters. Metadata ensures that data are interpretable, reliable, and relevant. Imagine a researcher accessing an imaging dataset without knowing the acquisition parameters —such ambiguity could lead to erroneous conclusions. By capturing metadata, we enhance data quality, facilitate cross-domain integration, and empower precision medicine approaches [3].

In summary, the harmonious interplay of standardized data and comprehensive metadata is the cornerstone of progress in health imaging research. In this direction, five EU projects (PRIMAGE, ChAImeleon, ProCAncer-I, INCISIVE, EuCanImage) worked together under the AI4HI (AI for Health Imaging) initiative [4], sharing experience and good practices towards the development of big data infrastructures [5] that will enable European, ethical and GDPR-compliant, quality-controlled, properly de-identified [6][7], cancer-related, medical imaging, and

other contextual clinical data platforms, in which both large-scale data and AI algorithms will co-exist. From the inception of those projects, it was clear that appropriate data and metadata models had to be devised to enable multicenter data harmonization and integration [8] and the development of AI models on top of them.

In this paper, we present the approaches on three of those projects, namely the ProCAncer-I, the ChAImeleon, and the INCISIVE projects explaining the rationale behind those projects, showing the methodology used for data collection and standardization, and reporting experiences and problems. Although a high-level paper already reported the initial design of all projects [8], in this chapter we repost the final implementation, providing experiences, problems, and solutions. Those approaches could be used as a starting point for new projects that aspire to enable the collection of large imaging datasets.

The remainder of this Chapter is structured as follows. In Section 2, we present the approach of the ProCAncer-I project, and in Sections 3 and 4 the approaches followed by the ChAImeleon and INCISIVE projects respectively. Then in Section 5, we discuss and compare the various approaches, whereas in Section 6, we conclude the chapter and provide directions for future research.

2. The ProCAncer-I project

2.1 Project description

In Europe, prostate cancer (PCa) is the second most frequent type of cancer in men and the third most lethal. Current clinical practices lead to overdiagnosis and overtreatment, necessitating more effective tools for discriminating between aggressive and non-aggressive diseases. The EU-funded ProCAncer-I project [9][10][11] proposes to develop advanced artificial intelligence models to address unmet clinical needs: diagnosis, metastases detection, and prediction of response to treatment. To achieve this, partners will generate a large interoperable repository of health images, and a scalable high-performance computing platform hosting the largest collection of PCa Magnetic Resonance Images used for developing robust PCa AI models. To ensure the rapid clinical implementation of the models developed, the project's partners will robustly monitor performance, accuracy, and reproducibility.

2.2 Data Collection & Standardization

The ProCAncer-I platform is designed to collect and manage large amounts of multimodal data and metadata to train advanced AI models for efficient and clinically oriented prostate cancer management. The platform storage, ProstateNet as shown in Figure 1, consists of three components: the DICOM Object Store, which stores medical imaging data; the Clinical Data Document Store, which stores clinical data; and the Meta-data Catalog, which stores metadata and semantic annotations to enable rich search and discovery of data and its exploitation.

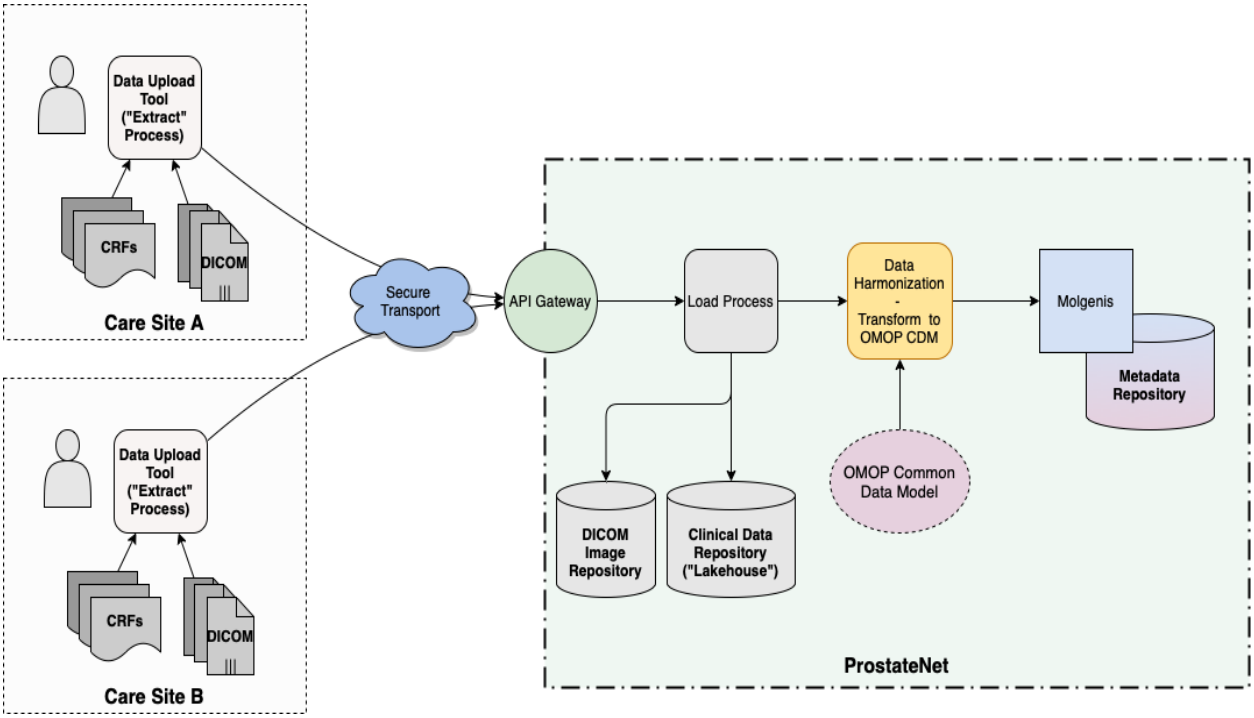

Figure 1. ProstateNet components and workflow.

Clinical partners use a local electronic case report form (eCRF) and data upload tool to organize DICOM studies, complete clinical information, validate use cases, anonymize data, and upload data to the cloud staging environment. In order to upload the data a first quality control is performed by the data upload tool, enabling only the uploading of cases that respect the minimum quality criteria, i.e. that the specified mandatory fields have been successfully filled out. Each clinical partner has its staging area. In the staging area, the users can run data curation tools, verify anonymization and quality/completeness of data, and submit validated cases to the ProstateNet repository.

In order to model the clinical data a common data model (CDM) is required that will allow the data homogenization and semantic integration of data that are uploaded by different clinical centers. The ProCAncer-I project uses the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [12] for harmonizing the clinical data. OMOP-CDM is one of the most widely used CDMs on a global scale, enabling large-scale data analysis in a consistent way. However, OMOP-CDM in its original form is primarily designed for general healthcare data, and its standard schema does not fully address the complexities and nuances of oncology data. The main reasons are that the oncology domain involves highly specific diagnoses that include cancer type, stage, grade, histology, and molecular characteristics, which the standard OMOP-CDM fields fail to adequately capture at this level of detail. In addition, cancer treatments are complex and often include combinations of surgery, radiation, targeted therapy, and hormone therapy. Capturing the specifics of such treatments as well as the ability to track disease progression, response to treatment, and recurrence over time is information that has not been adequately represented in the standard OMOP-CDM. For the above reasons, ProCAncer-I also employs the OMOP Oncology standard CDM extension [13],

to represent cancer data at the levels of granularity and abstraction required to support cancer research. In addition, the Oncology extension incorporates standards from oncology-specific ontologies and classifications (e.g., ICD-O, NAACCR, and NCIT), through the OMOP cancer modifiers, ensuring consistency and interoperability with other data standards and improving the ability to conduct multi-institutional studies in the oncology domain.

While electronic phenotyping for creating cohort populations has been extensively studied and implemented within the OHDSI community, building cohorts that combine clinical data with imaging metadata remains a challenge. This difficulty arises due to the lack of a comprehensive standard for image-related information and the necessary image curation processes within the OMOP-CDM. These curation processes include data quality measures like motion correction and co-registration, as well as important annotation and labeling tasks, such as identifying anatomical structures or noting specific pathologies with segmentation masks.

Although the Digital Imaging and Communications in Medicine (DICOM) standard is used for medical imaging data—and can provide access to critical image acquisition parameters (e.g., acquisition method, the field of view, slice thickness) for cohort discovery through its DICOM tags, —it unfortunately falls short in some areas. Specifically, DICOM lacks standardized information crucial for identifying relevant images. For instance, whether an MR series is a T2-weighted axial series is typically recorded in the DICOM tag Series Description (0008,103E), which is free text and can vary widely between clinical institutions.

To address these challenges, the ProCAncer-I project introduced a Medical Imaging extension, the MI-CDM [14][15] that focuses on incorporating standardized imaging metadata for the key DICOM tags used for cohort discovery and extends the model to include image curation processes. This enables the retrieval of curated images based on standardized vocabularies and maintains the connection between original and processed images for tracking provenance. RadLex [16] is used as a reference ontology for standardizing the radiology-related metadata and harmonizing diverse imaging methods, leading to a more complete understanding of patient health, and improving diagnostic precision and personalized treatment strategies by integrating both clinical and radiological data into decision models. An example of the MI-CDM instantiation is shown in Figure 2.

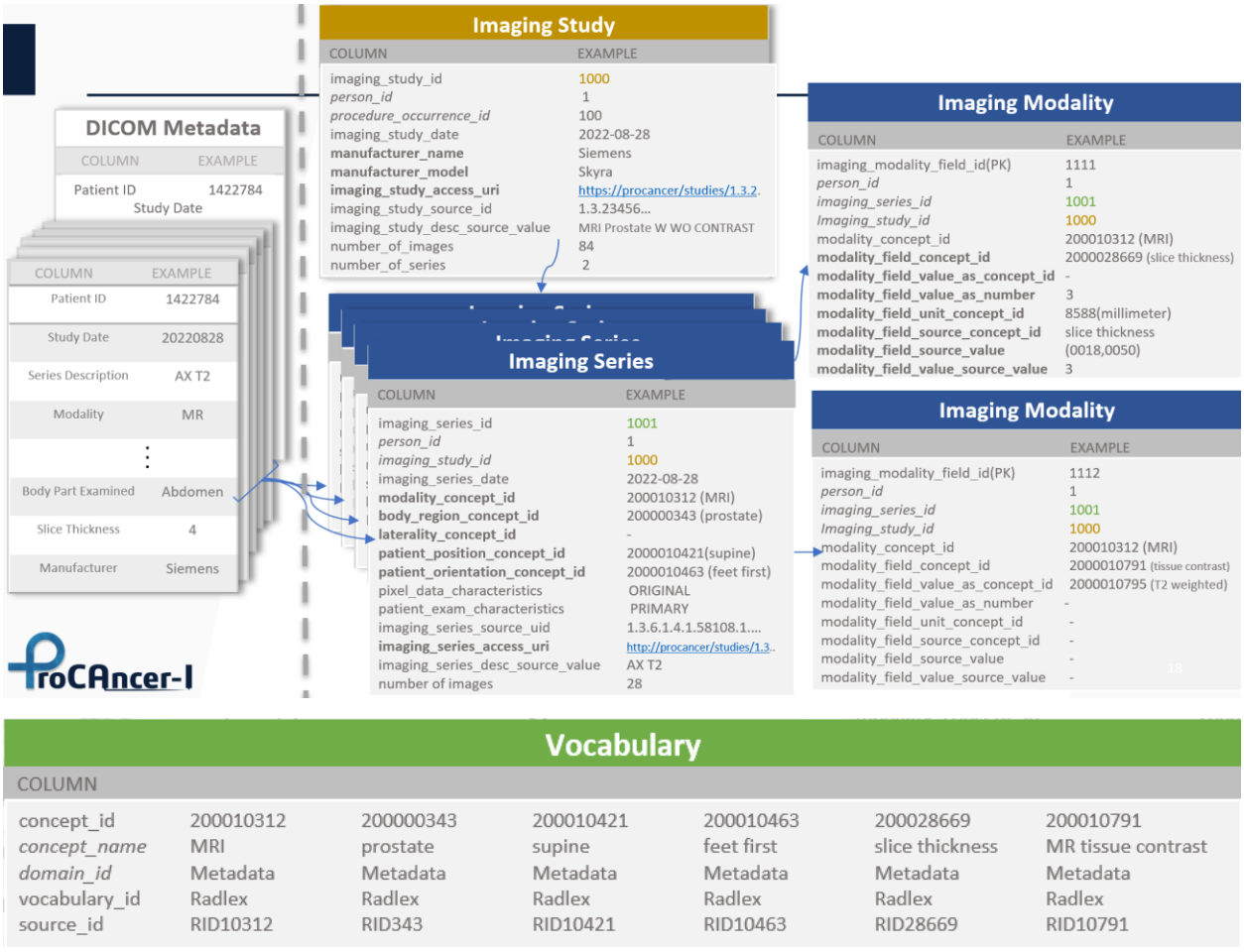

Figure 2: Example of the MI-CDM imaging metadata model instantiation.

Finally, having a common data model to represent and store data, a meta-data catalog can facilitate searching the data collection. The ProCAncer-I project uses the MOLGENIS platform [17] as the primary metadata catalog.

2.3 Experiences and problems

The MI-CDM extension is a significant advancement in the integration of imaging data in the OMOP-CDM, but it has some limitations. First, the extension relies on the existing DICOM standard, which lacks standardized information, necessitating careful mappings and integration efforts that may vary across institutions and settings. For example, even though the MI-CDM enables the storage of important acquisition parameters in a standardized format through the Radlex vocabulary, extracting this information from the DICOM tag requires the development of an AI model that can identify the sequence type automatically. Second, the extension was primarily guided by the demands of prostate cancer imaging within the ProCAncer-I project, therefore all standardized vocabularies are limited to the use cases/cohorts that the project needs to address. Extending its use to other medical areas would require validation, potentially introducing complexities in standardizing specific attributes and metadata. Finally, the MI-CDM's success depends on the adoption and standardization of its terminologies within the broader

healthcare and research community. Collaborative efforts are essential to ensure consistent implementation, harmonization with existing standards, and the continued evolution of the MI-CDM framework.

3. The ChAImeleon project

3.1 Project description

The ChAImeleon [18] project is part of the Horizon 2020 funding program, and its primary objective is to establish a pan-European Distributed Data Repository with quality-checked imaging and clinical data, that fully complies with legal and ethical regulations, facilitating open reuse in Artificial Intelligence (AI) experimentation for cancer management. This repository is populated with anonymized multimodality imaging and in some cases, with related clinical data for diagnosed patients of lung, prostate, breast, and colorectal cancer. Also, the project aims to be interoperable with existing repositories, enabling secure sharing and reuse of anonymized imaging and clinical data, for AI developers to develop cancer management tools and solutions.

The ChAImeleon project actively engages with Real World Data (RWD), which involves collecting data through the routine delivery of healthcare without specific enrollment conditions. Initially designed to incorporate data from 10 partners across 10 European countries, the project faces considerable heterogeneity in the data. To effectively meet the project's objectives and ensure interoperability, the adoption of standardized approaches becomes imperative. Interoperability, in this context, entails not only working with existing standards but also actively contributing to their development, all while adhering to the FAIR [19] (Findable, Accessible, Interoperable, and Reusable) data principles.

3.2 Data Collection & Standardization

The main Data Standards initially proposed for the project were:

- For Medical Imaging, the DICOM standard is proposed due to its widespread adoption in the visualization and sharing of medical images and associated information [20]. In clinical practice, a medical image is a DICOM file that includes a header in addition to the actual image [21][22]. The header stores information categorized into groups of elements known as "DICOM tags." These tags, in turn, represent real-world entities such as images, procedures, or interpretation reports within the DICOM semantic data model, utilizing templates of attributes or data elements, with each tag identifying a specific attribute.
- Regarding Clinical Data, the Common Data Model (CDM) selected for the project was OMOP (Observational Medical Outcomes Partnership) [12]. The main reasons were that it is an international reach; it was built for the analysis of observational health data; covers data protection by design; is a patient-centered model; is open-source and does not require specific technology for its use; is scalable (optimized for data sources varying in size up to hundreds of billions of patients and billions of clinical observations) and extensible (several

extensions available to cover all the types of clinical data); and last but not least, OMOP benefits from a community that actively contributes to its development.

The most relevant clinical data related to the diagnosis, initial treatment, and first follow-up of the disease were selected. Regarding medical images, mandatory modalities were defined for each type of cancer, but optionally, the upload of other images related to the disease was allowed.

Data collection in ChAImeleon took place internally in each data provider center, and each center adopted its own strategy based on its characteristics. Medical images were extracted directly from the PACS systems of each center in DICOM format, and prepared for export to the central repository. The harmonization of the images took place in the central platform.

Clinical Data, in most cases, was manually entered into a local electronic Case Report Form (eCRF) created specifically for the project. In some cases, the data was already in OMOP format, so cases and variables were automatically selected and ingested automatically into the local eCRF. Another approach was to directly complete clinical data in an Excel sheet, which was prepared for subsequent automatic ingestion into the local eCRF. In all cases, once the clinical data was validated on the local platform, it was fully anonymized and sent to the central platform.

3.3 Experiences and problems

In medical imaging, the DICOM standard is widely adopted and accepted by the scientific community for both primary and secondary use. In this regard, there have been no issues in adopting this standard. However, when it comes to clinical data, the use of OMOP CDM has caused more discussion. The adoption of the OMOP CDM has caused significant challenges in several aspects.

For example, when dealing with real-world data, we face situations where it is very difficult to obtain some information, e.g., for certain procedures performed in other centers or hospitals. In the case of a patient receiving treatment in a hospital other than his or her primary care center, it may not be possible to determine the date of treatment. This lack of date information poses a problem of non-compliance with the OMOP-CDM because, in OMOP, those dates are mandatory.

Similarly, the OMOP CDM lacks the ability to interpret data relating to non-existent entities. For example, if physicians determine that a patient has no history of cancer, translating this information into the OMOP CDM framework is complicated. This raises limitations to explicitly save and differentiate missing and negative data.

Currently, OMOP does not cover information stored in DICOM metadata. Information essential for radiological research studies, such as the acquisition parameters used for a particular study or the anatomical regions examined, is not easily accommodated in CDM. In addition, when the image is processed, the results of the procedure are not easy to store either, for instance, radiomics feature extractions.

Furthermore, comprehensive characterization of cancer necessitates more detailed information than what is typically required for an observational study [23]. Details such as disease staging and grading, metastatic spread, and information about treatments administered to patients (e.g., chemotherapy, radiotherapy, hormonotherapy, or combinations thereof) are essential. Additionally, the patient journey requires specific terminology that is often not specified in the source document but is highly relevant, such as overall survival, progression-free survival, adjuvant period, or disease recurrence. The OMOP CDM is not ready to support this kind of information. Due to the issues raised in the previous section, it became imperative to incorporate both the OMOP medical imaging extension and the OMOP oncology extension to comprehensively encompass all data relevant to these fields.

Oncology extension. The oncology extension is currently integrated into OMOP CDM version 5.4 and is planned to be fully integrated with CDM in the upcoming version 6.1 [13]. The main feature of this extension is the creation of two new tables 'episode' and 'episode_event'. The episode table aggregates lower-level clinical events (visit_occurrence, drug_exposure, procedure_occurrence, device_exposure) into a higher-level abstraction representing clinically and analytically relevant disease phases, outcomes, and treatments. The episode_event table connects qualifying clinical events (visit_occurrence, drug_exposure, procedure_occurrence, device_exposure) to the appropriate episode entry. For example, cancers including their development over time, their treatment, and final resolution. The episode_event table connects qualifying clinical events (such as condition_occurrence, drug_exposure, procedure_occurrence, and measurement) to the appropriate episode entry. For example, linking the precise location of the metastasis (cancer modifier in measurement) to the disease episode. The measurement table contains records of Measurement, i.e., structured values (numerical or categorical) obtained through systematic and standardized examination or testing of a Person or Person's sample. The measurement table contains both orders and results of such Measurements as laboratory tests, vital signs, quantitative findings from pathology reports, etc. Measurements are stored as attribute-value pairs, with the attribute as the Measurement Concept and the value representing the result. The value can be a Concept (stored in value_as_concept), or a numerical value (value_as_number) with a Unit (unit_concept_id).

Medical Imaging Extension. To address the limitations when dealing with imaging data, a pilot implementation of a new Radiology table, based on RadLex standardized vocabulary for the radiology process [24] was initially proposed. Later on, OHDSI proposed an extension as illustrated in Figure 4.

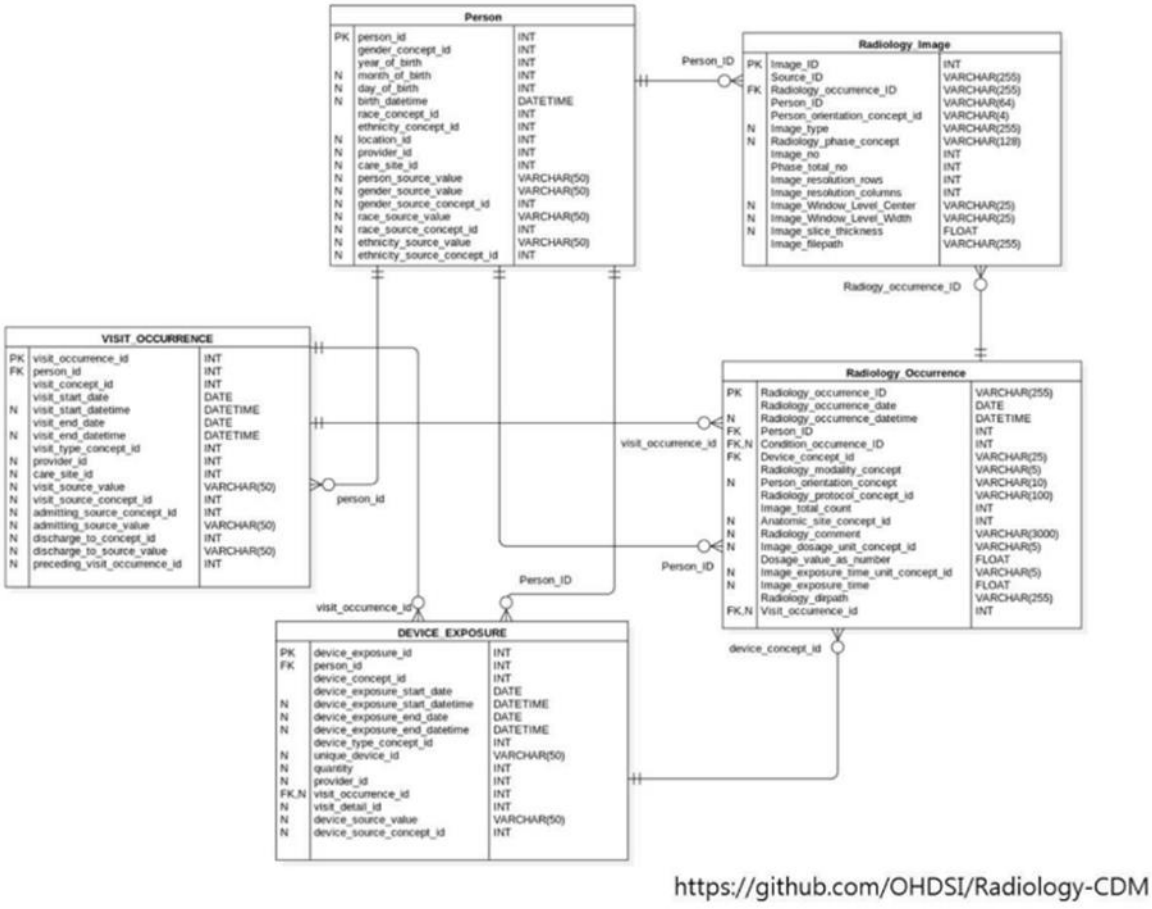

Figure 3. Proposed Radiology extension to OMOP

However, this proposed extension lacks key acquisition parameters such as slice thickness, bolus/contrast agent usage, and MRI acquisition details like T1, and T2 information, diffusion coefficients, b-values, magnetic field strength, etc. Given the significance of these parameters in radiological research studies, there exists an opportunity for the CHAIMELEON project to spearhead standardization efforts regarding OMOP terminologies within the OMOP framework. Additionally, there is potential for the project to develop new ontologies pertaining to quantitative radiological features. In 2022, the official medical imaging working group of OHDSI was created to formalize the extension and address all the issues seen in the previous proposal. At that time, CHAIMELEON joined the working group to create synergies between the two groups and facilitate the creation of the new extension. One of the first outcomes of this working group was the creation of a poster titled 'Development of the Medical Imaging Extension' [26], where the extension proposal was presented to the OHDSI community, along with a use case. This poster concludes with an intention statement for the imaging medical extension: "We propose a medical imaging extension to standardize features and provenance of medical images in OMOP-CDM. With further development, we hope that the medical image extension provides the essential infrastructure for robust, scalable, and reproducible medical image study."

In September 2023, the first publication of the extension was released [24] [25], where the proposal is presented in more detail. The most relevant changes include the creation of two

tables (Figure 4). The Image_occurrence table describes imaging events and provides data lineage to the imaging study. The Image_occurrence table has three functions. First, the Image_occurrence table links to DICOM images at a study or a series level; Second, the Image_occurrence table includes series-level parameters. Third, the Image_occurrence table provides provenance for the Image_feature table to identify the images used in creating the features. The Image_feature table describes the characteristics of the images and their provenance. Each row will contain a uniquely identified feature with links to the source imaging as well as the clinical domain table the feature is located in.

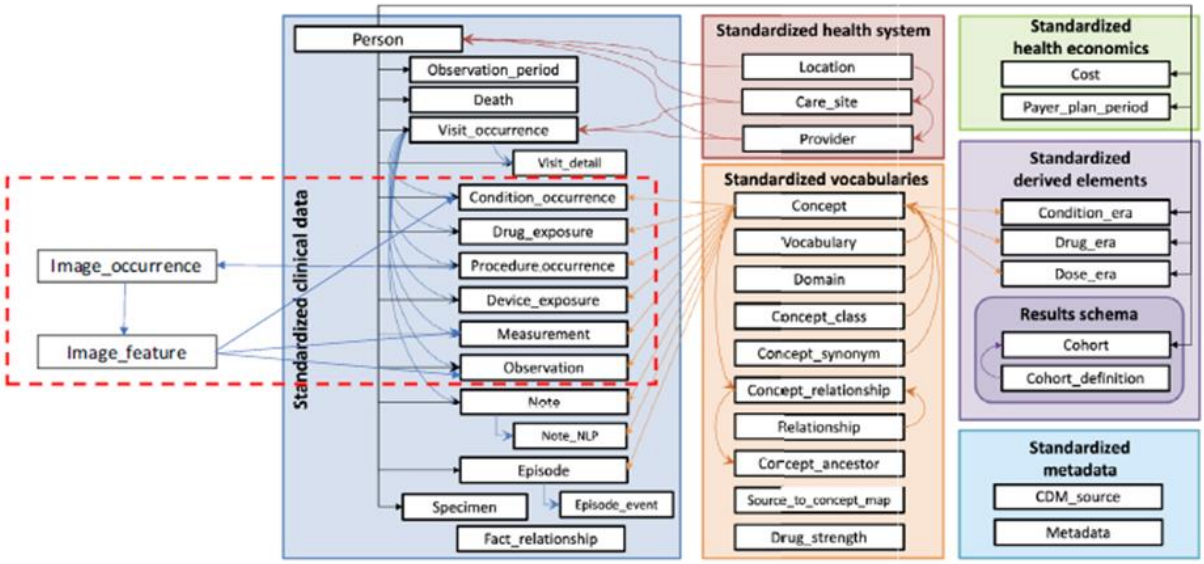

Figure 4. Incorporation of the proposed image data model to existing OMOP CDM v5.4

Final Data Standard proposal. When implementing the oncology extension in CHAIMELEON, we realized that our project was not capturing the complete patient journey. Instead, we are centered on imaging data, and we were taking a 'snapshot' of a specific moment, which is their diagnosis and initial treatment, with a subsequent defined follow-up period (no more than 2 years, depending on the tumor type), where we studied their survival and recurrence. In this scenario, where it is not as crucial to record the different phases of the patient's evolution such as the (neo)adjuvant period, or subsequent lines of treatment, it was deemed that the effort required to adopt those tables did not justify the relevance of the information to be collected. For that reason, it was decided not to implement the ‘episode’ and ‘episode_event’ tables.

On the other hand, it was deemed relevant to capture all the specific descriptions provided by this complex disease. Therefore, it was decided to adopt the use of recommended vocabularies by ODHSI [27]: SNOMED, ICDO3, HemOnc, RxNorm, NAACCR, and Episode Type.

Furthermore, there were certain concepts that were not adequately captured in the ontologies. Hence, the decision was made to utilize [27] the 'custom vocabulary' offered by the OHDSI community in those instances for the fast implementation approach (it is recommended to expand the 'concept' table starting from over 2 billion). For a long-term approach, it was decided to contact OHDSI in order to include these terms in the vocabulary [27].

Regarding the medical imaging extension, waiting for the formalization of the new extension was not a viable option due to the need to populate the database as soon as possible within the project's described timeline, and using the initial proposal did not address all the project's needs.

For all those reasons, the adoption of the OMOP CDM would lead to loss of data. So, it was decided to use OMOP as much as possible and complement it through the use of XML and JSON files sharing the OMOP structure. So far, a hybrid structure has been defined, based on the eCRF's structure completed with OMOP information.

4. The INCISIVE project

4.1 Project description

INCISIVE, a 42-month EU-funded project [28], has a dual aim: along with the generation of an AI toolbox that will provide decision-making assistance to daily healthcare practice, the end goal of the INCISIVE is the implementation of a pan-European repository of health images following a federated approach. In that sense, INCISIVE is called to address many challenges in terms of data availability, data harmonization, and data integration. The main focus is the aggregation and homogenization of multisite clinical and imaging data that come from multiple healthcare centers across Europe, towards the creation of a homogenized distributed repository that will enable Artificial Intelligence (AI) development, and, consequently, the adoption of AI solutions in health imaging. In this respect, data harmonization constitutes a major pillar.

4.2 Data Collection & Standardization

Data collected within the INCISIVE studies are prepared following a protocol established within the project and are aligned with the project's needs and requirements. Data pertain to four of the most common cancer types: breast, colorectal, lung, and prostate. Data consists of two major data types, clinical metadata and imaging data. The clinical metadata is provided through a well-defined and structured template in the '.xls' format, filled in with information extracted from the healthcare center's Electronic health record system. This template contains information referring to (i) patient medical history, (ii) diagnosis, (iii) follow-up examinations at specific time points, (iv) histopathological examination results, and (iv) laboratory examination results. Regarding the imaging data, multiple modalities are collected for each patient. The type of imaging modalities collected are different between the four cancer types, and the requirements for what type of modalities are provided for each case were defined in the study protocol. Imagining data are provided de-identified in DICOM format, along with the segmentation and annotation files in nifty format. The initial design of the study has established a connection between the clinical metadata and the de-identified imaging data through a unique patient codification.

The diversity of data that originates from multiple sources leads to the need for data harmonization and data standardization is the key to allowing this unification of data. The approach followed in the project for data standardization consists of five steps referring to

structural, semantic, and syntactic interoperability. These steps contribute to the ultimate goal of this procedure, the creation of a common data model (CDM) that all data should comply with. The four steps of this procedure, referring to the clinical metadata alone are: (i) definition of the clinical metadata and structure, (ii) Content standardization, (iii) Attribute standardization, (iv) Syntactic standardization, and, (v) transformation. Each one of the steps is described in detail below, along with the challenges faced in the process and depicted in Figure 1.

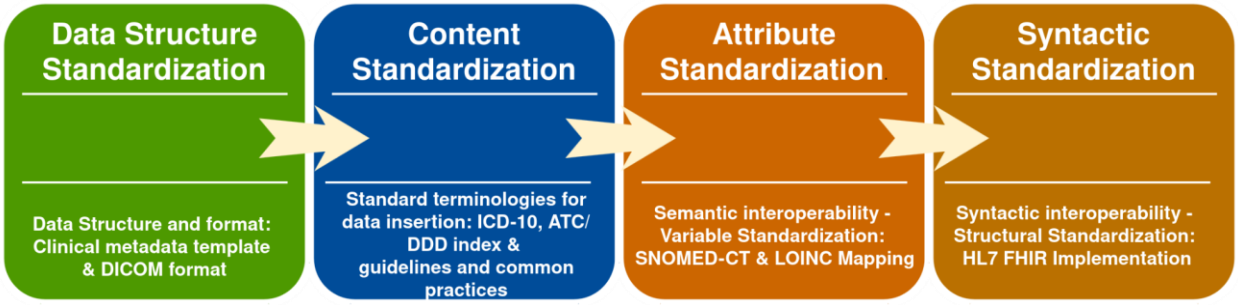

Figure 1. The INCISIVE Data Standardization Approach

Data Structure Standardization. Referring to the data structure, two approaches for standardization were followed. The first one, regarding the clinical metadata that accompanies the imaging data holding information on the patient status, refers to an iterative procedure that took place in order to decide the manner in which this kind of data will be provided. A set of clinical variables was pre-selected by the technical team based on the needs and requirements of the project, bibliography, and prior knowledge. This set of variables was circulated through the clinical partners, discussed through workshops, and finally resulted in the final list of variables that will be collected along with the imaging data. This set of variables was cataloged in an Excel file as different sheets referring to Demographic information, Diagnosis, Follow-ups, Treatment, Histopathological Information, and Laboratory Examinations. The second approach referring to imaging data, was to follow a well-defined imaging examination protocol, the DICOM standard, avoiding any inconsistency that may arise in the procedure of extracting the data from the institutes' data management systems, and specifically in Picture Archiving and Communication System (PACS) systems.

Content Standardization. This step includes the standardization of the templates' content, in terms of the actual data inserted in the Excel file, and the standardization of labels used in the annotation procedure for the characterization of the findings. In the first case, terminologies based on medical standards and best practices were adopted. Specific standards were proposed to achieve the data homogeneity: (i) ATC/DDD Index 2024 codes for medications, (ii) ICD-10 for cancer classification and tumor characterization, and (3) consensus decisions based on prior experience and international guidelines. By utilizing these standards for the terminology used in clinical metadata provision, the project ensured the content homogenization of the repository. During the design procedure, all medical partners were asked to provide sample data.

Attribute Standardization. This step focuses on the semantic interoperability of data. Semantic interoperability refers to the ability of different systems to exchange information with a common

understanding. The main goal of this step is to ensure that the provided data from various sites is not only structurally compatible (following the same format or structure) but also semantically compatible. After the completion of the two previous steps, and structural and content standardization completion, the next step is to ensure semantic interoperability by standardizing the actual fields of the template. This was done by performing a complete mapping of each field to two well-established standards for medical information, SNOMED-CT (https://www.snomed.org/) and LOINC (https://loinc.org/). The former was used to convert the field names referring to clinical, diagnostic, and imaging-related data into specific concepts using the coding system provided by the standard, and the latter was used in the same way but for laboratory data.

Syntactic Standardization. This step focuses on the syntactic interoperability of data. Syntactic interoperability is the capability of various systems to share and interpret data accurately, relying on a shared syntax or structure. It is achieved by ensuring that the formatting and structure of data follow standardized practices across a range of systems. Thus, in this step, the HL7 FHIR [29] standard was used, which is used to organize data under specific definitions, called resources. Resources share a common way of defining and representing them, building them from data types that define common reusable patterns of elements, a common set of metadata , and a human-readable part. To that end, this step includes the procedure where (i) the proper FHIR resources to represent the data were identified, (ii) the variables were mapped to specific attributes of each FHIR resource, and, (iii) the structure of the FHIR resource was implemented in a file in FHIR XML format. Since the HL7 FHIR protocol is a complete and well-established protocol, no challenges were faced at this step. However, this standard does not yet fully support the incorporation of imaging data in the CDM, so instead of including the imaging metadata information in the CDM and in the FHIR server, Imaging examinations were separately stored in a PACS server, preserving all the information needed and the connection with clinical metadata.

Extraction/Transformation/Load. The final step is the implementation of an Extraction, Transformation, and Loading (ETL) process that allows data uploaded to the repository following the predefined structure to be transformed into the CDM's standards. This process executes these three basic steps: (i) extraction of data that is provided to an .xls format, (ii) transformation of this data into FHIR xml messages, following the protocols described above, and, (iii) loading of data into the FHIR server.

4.3 Experiences and problems

During this procedure, some challenges were encountered that needed to be addressed in order to overcome the problems and result in a solid result. In the step of data structure standardization, some inconsistencies were identified in the way that data are collected and stored among the various sites participating in the project, but through this procedure, the consensus of variables to be included in the study was unified. Through the examination of data in the content standardization step, many inconsistencies were identified in terms of values provided. Each organization uses different protocols to store the data, resulting in many

variations between the provided datasets. Some characteristic examples are the units used for blood examination metrics, the way that staging is reported (e.g. TNM or overall staging), or the way that treatment is declared (e.g. drug commercial name against substance). Through this content standardization procedure, all data collectors should follow the same protocols and provide data in a unified, consistent, and homogenized way. Referring to the standardization of annotation labels, a common vocabulary was defined to be used through the segmentation and annotation procedure by all data collectors. During the third step of attribute standardization, some challenges were faced concerning the use of both standards. In the mapping of laboratory examinations to the LOINC standard, the choice of the proper codification was not always clear, thus in many cases, consultancy with the medical partners was needed in order to decide the proper codification of some variables. In the mapping of the rest of the variables, there were some cases where the codification was not possible due to a lack of direct mapping of a field to a SNOMED-CT code. However, the specific standard is internationally widely used, with an active development and support team, and provides the possibility of dynamically adding new terms. In terms of Syntactic standardization, the problem that arose was that the standard used, HL7 FHIR, does not support the incorporation of imaging data, so a PACS server needed to be employed in parallel with the FHIR server.

Following this procedure and by addressing the issues raised during the process, the INCISIVE project achieved homogenized data coming from different sources into a format in which the data share the same principles achieving the two main goals: (i) enabling the implementation of a well-structured federated repository through the provision of data availability and search information and (ii) facilitating data analysis and the development of robust and accurate AI services.

5. Discussion

Table 1 summarizes the approaches of the three projects reported in this paper. As shown, all projects focus on cancer, with ChAImeleon and INCISIVE focusing on several types whereas the ProCAncer-I project focuses on a single cancer, i.e. prostate cancer. Further, all projects collect imaging data along with the corresponding clinical metadata, whereas INCISIVE collects in addition biological metadata.

Regarding models used for imaging data, all projects use the DICOM standard, whereas clinical data resolves to bottom-up models (INCISIVE) designed by collecting the relevant fields from clinicians and using multiple standard terminologies for collecting data through eCRFs. On the other hand, both ProCAncer-I and ChAImeleon adopt OMOP-CDM, however the model, although prominent in the field, fails to model all relevant information required by the projects. As such both projects adopt extensions (i.e. the oncology extension) and eventually they reside to their own extension for capturing all relevant information required.

Curation tools rely on the selected models in order to perform a quality check on the ingested data and to enable a search of the data based on their metadata.

Table 1. Summary of the approach of the three projects.

	ProCAncer-I	ChAImeleon	INCISIVE
Domain	Prostate Cancer	Lung, breast, prostate, and colorectal cancer	Lung, breast, colorectal, and prostate cancer
Collected Data	Imaging / Clinical Metadata	Imaging / Clinical Metadata	Imaging / Clinical Metadata Biological Metadata
Existing Models Used	DICOM Radiation Therapy for imaging data OMOP-CDM for clinical data Oncology CDM extension RadLex for harmonizing diverse imaging methods and metadata	DICOM for imaging metadata OMOP-CDM for clinical OMOP medical imaging extension Oncology CDM extension	DICOM for imaging metadata Multiple terminologies for clinical data (e.g., SNOMED-CT, LOINC, ICD10, ATC classification) FHIR for communication
Extensions Provided	MI-CDM: an extension to OMOP-CDM going beyond radiology/oncology extensions		Designed a Common Data Model based on commonly agreed fields between clinical centers
Model & MetaData Tools Used	MOLGENIS platform as the primary metadata catalog	The CHAIMELEON repository allows search using the selected metadata model	Data Integration Quality Check Tool employed to identify whether data follow the harmonization requirements defined

Essentially, as per the experiences of those three projects, no data model will cover all the details each individual project requires. DICOM although useful, requires standardized metadata which is not available out of the box. On the other hand, OMOP-CDM is not able to appropriately describe yet the imaging datasets and their curation for the deployment of AI models on top. Those directions should be explored in the near future and require extensions. However, the optimal way to do this is not through ad-hoc solutions but through extending the community standards through joint standardization efforts that have the potential to have a wider impact on the community.

6. Conclusion

This paper presents an overview of the standardization efforts implemented within three EU projects focusing on cancer focusing on establishing a common data model for imaging and

clinical data and storing appropriate metadata along with the imaging data. Being able to homogenously access data that might have been captured from different hospitals has the potential to improve patient care and reduce costs. With the ability to harness and analyze large volumes of data, healthcare can be revolutionized, unlocking new insights, improving patient safety, and enhancing operational efficiency.

References

[1] Brancato, V., Esposito, G., Coppola, L. et al. Standardizing digital biobanks: integrating imaging, genomic, and clinical data for precision medicine. J Transl Med 2024;22:136. https://doi.org/10.1186/s12967-024-04891-8.

[2] Lekadir, K., Osuala, R., Gallin, C., et al. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv preprint 2023. arXiv:2109.09658.

[3] Badawy, R., Hameed, F., Bataille, L., Little, M. A., Claes, K., Saria, S., ... & Karlin, D. R. Metadata concepts for advancing the use of digital health technologies in clinical research. Digital Biomarkers 2020;3(3):116-132.

[4] AI for Health Imaging. Available Online: https://ai4hi.net/, visited April 2024.

[5] Kondylakis, H., Kalokyri, V., Sfakianakis, S. et al. Data infrastructures for AI in medical imaging: a report on the experiences of five EU projects. Eur Radiol Exp 2023;7:20. https://doi.org/10.1186/s41747-023-00336-x

[6] Clunie, D., Prior, F., Rutherford, M., et al. Summary of the National Cancer Institute 2023 Virtual Workshop on Medical Image De-identification-Part 1: Report of the MIDI Task Group - Best Practices and Recommendations, Tools for Conventional Approaches to De-identification, International Approaches to De-identification, and Industry Panel on Image De-identification. J Imaging Inform Med. 2024 Jul 12. doi: 10.1007/s10278-024-01182-y. Epub ahead of print. PMID: 38997571.

[7] Kondylakis, H., Catalan, R., Alabart, S.M., et al. Documenting the de-identification process of clinical and imaging data for AI for health imaging projects. Insights Imaging. 2024 May 31;15(1):130. doi: 10.1186/s13244-024-01711-x. PMID: 38816658; PMCID: PMC11139818.

[8] Kondylakis, H., Ciarrocchi, E., Cerda-Alberich, L., Chouvarda, I., Fromont, L. A., Garcia-Aznar, et al. & AI4HealthImaging Working Group on metadata models**. Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks. European Radiology Experimental 2022;6(1):29.

[9] Kalokyri, V., Tachos, N., Sfakianakis, S., et al. Data preparation for artificial intelligence in medical imaging: Experiences from the ProCAncer-I initiative. IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biolog, 2023, pp. 73-74. IEEE.

[10] Colantonio, S., Berti, A., Buongiorno, et al. AI trustworthiness in prostate cancer imaging: a look at algorithmic and system transparency. In 2023 IEEE EMBS Special Topic Conference on Data Science and Engineering in Healthcare, Medicine and Biology (pp. 79-80). IEEE.

[11] Kondylakis, H., Sfakianakis, S., Kalokyri, V., Tachos, N., Fotiadis, D., Marias, K., & Tsiknakis, M. Data ingestion for AI in prostate cancer. Challenges of Trustable AI and Added-Value on Health, 2022:244-248.
[12] OMOP-CDM. Available Online: https://www.ohdsi.org/data-standardization/, visited April 2024.
[13] OMOP-CDM Oncology Extension. Available Online: https://ohdsi.github.io/CommonDataModel/oncology.html, visited April 2024.
[14] Kalokyri, V., Kondylakis, H., Sfakianakis, et al. MI-Common Data Model: Extending Observational Medical Outcomes Partnership-Common Data Model (OMOP-CDM) for Registering Medical Imaging Metadata and Subsequent Curation Processes. JCO Clinical Cancer Informatics 2023;7:e2300101.
[15] Kalokyri, V., Kondylakis, H., Sfakianakis, et al. R-CDM: Extending OMOP-CDM for registering tomography imaging. European OHDSI Symposium, 2023
[16] RADLEX radiology lexicon. Available Online: https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon, visited April 2024.
[17] MOLGENIS. Available Online: https://www.molgenis.org/, visited April 2024.
[18] ChAImeleon project. Available Online: https://chaimeleon.eu/, visited April 2024.
[19] FAIR Principles. Available Online: https://www.go-fair.org/fair-principles/, visited April 2024.
[20] Bidgood, W.D. Jr, Horii, S.C., Prior, F.W., Van Syckle, D.E. Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging. J Am Med Inform Assoc.1997;4(3):199-212.
[21] Bidgood, W.D., Horii, S.C. Introduction to the ACR-NEMA DICOM standard. RadioGraphics. 1992;12(2):345-55.
[22] Mildenberger, P., Eichelberg, M., Martin, E. Introduction to the DICOM standard. Eur Radiol. 2022;12(4):920-7.
[23] Belenkaya, R., Gurley, M.J., Golozar, A., Dymshyts, D., Miller, R.T., Williams, A.E., et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. JCO Clin Cancer Inform 2021:12-20.
[24] Park, C., You, S.C., Jeon, H., Jeong, C.W., Choi, J.W., Park, R.W. Development and Validation of the Radiology Common Data Model (R-CDM) for the International Standardization of Medical Imaging Data. Yonsei Med J. 2022;63(Suppl):S74-83.
[25] Park, W.Y., Jeon, K., Schmidt, T.S., et al. Development of Medical Imaging Data Standardization for Imaging-Based Observational Research: OMOP Common Data Model Extension. J Imaging Inform Med. 2024 Apr;37(2):899-908.
[26] Development of the Medical Imaging Extension for OMOP-CDM. Available Online: https://www.ohdsi.org/2022showcase-26/, visited April 2024.
[27] New concepts for Cancer modifier vocabulary. Available Online: https://github.com/OHDSI/Vocabulary-v5.0/issues/551, visited April 2024.
[28] INCISIVE project. Available Online: https://incisive-project.eu/, visited April 2024.
[29] HL7 FHIR. Available Online: https://www.hl7.org/fhir/index.html, , visited April 2024.