Data and System Traceability for Transparent AI in Medical Imaging

Sara Colantonio[1,*], Andrea Berti[1,2], Gianluca Carloni[1,2], Claudia Caudai[1], Giulio Del Corso[1], Danila Germanese[1], Eva Pachetti[1,2], Maria Antonietta Pascali[1], Varvara Kalokyri[3], Haridimos Kondylakis[3], Charalampos Kalantzopoulos[4], Nikolaos Tachos[4], Dimitris Fotiadis[4], Valentina Giannini[5,6], Simone Mazzetti[5,6], Daniele Regge[5,6], Nickolas Papanikolaou[7], Konstantinos Marias[3] and Manolis Tsiknakis[3]

[1]*Institute of Information Science and Technologies, National Research Council of Italy (ISTI-CNR), Pisa, Italy*
[2]*University of Pisa, Pisa, Italy*
[3]*Foundation for Research and Technology Hellas (FORTH), Institute of Computer Science, Heraklion, Greece*
[5]*Foundation for Research and Technology Hellas, Ioannina, Greece*
[5]*Department of Surgical Sciences, University of Turin, Turin, Italy*
[6]*Department of Radiology, Candiolo Cancer Institute, FPO-IRCCS, Candiolo, Italy*
[7]*Champalimaud Foundation, Computational Clinical Imaging Group, Lisboa, Portugal*

Abstract

Artificial intelligence holds the promise to revolutionize medical practices, particularly the realm of image-based diagnostics. Nonetheless, the integration of artificial intelligence technologies brings forth a range of immediate and future challenges that are the focus of almost all related discussions. A responsible approach to the development and use of artificial intelligence is essential to effectively address and mitigate these challenges, via strong scientific foundations, technical reliability, thorough testing and validation procedures, risk assessment and alignment with ethical principles. Central to this is the principle of transparency, as a key ingredient to foster trust and reliability. Transparency can be upheld through measures such as disclosing data sources and their use, as well as demonstrating transparent system development, operation and use. In this respect, it is strictly interconnected with the traceability of data and AI systems. This discussion paper briefly outlines the most relevant issues related to transparency and the methods used in the EU H2020 ProCAncer-I project to fulfill its mandates, in terms of data and system traceability, also linked to other projects, such as the Tuscany Region's NAVIGATOR project, and in compliance with the requirements of the FUTURE-AI guidelines.

Keywords

Transparent Artificial Intelligence, traceability, oncologic imaging, AI Model Passport

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy
*Corresponding author.
sara.colantonio@isti.cnr.it (S. Colantonio); andrea.berti@isti.cnr.itl (A. Berti); gianluca.carloni@isti.cnr.it (G. Carloni); claudia.caudai@isti.cnr.it (C. Caudai); giulio.delcorso@isti.cnr.it (G. D. Corso); danila.germanese@isti.cnr.it (D. Germanese); eva.pachetti@isti.cnr.it (E. Pachetti); maria.antonietta.pascali@isti.cnr.it (M. A. Pascali); vkalokyri@ics.forth.gr (V. Kalokyri); vkalokyri@ics.forth.gr (H. Kondylakis); xkalantzopoulos@gmail.com (C. Kalantzopoulos); ntachos@gmail.com (N. Tachos); fotiadis@cs.uoi.gr (D. Fotiadis); valentina.giannini@unito.it (V. Giannini); simone.mazzetti@ircc.it (S. Mazzetti); daniele.regge@ircc.it (D. Regge); nickolas.papanikolaou@research.fchampalimaud.org (N. Papanikolaou); kmarias@ics.forth.gr (K. Marias); tsiknaki@ics.forth.gr (M. Tsiknakis)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. Introduction

In the realm of clinical applications, the impact and adoption of Artificial Intelligence (AI) and Machine Learning (ML) technologies hinge on their ability to demonstrate reliability and clinical usefulness, ensure absolute patient safety, and earn the trust and approval of clinical end users and stakeholders [1]. However, trust is a dynamic and multi-layered concept that extends beyond technical performance to encompass psychological, sociological, philosophical, and ethical dimensions[1]. It is shaped by both objective factors and subjective perceptions and beliefs. Consequently, the extensive efforts by the scientific, regulatory, and standardization communities aimed so far to identify crucial elements AI systems must possess to address concerns and foster trust among users.

A critical element in building trust is a commitment to transparency. The High-Level Expert Group (HLEG) on Artificial Intelligence's guidelines [2] outlined transparency through three aspects: *traceability, explainability, and frank disclosure of an AI system's limitations.* The newly ratified AI Act [3] also emphasizes transparency as a core objective. It is designed to ensure individuals understand the design and usage of AI systems, along with the responsibilities companies and public authorities have concerning decisions made by AI. The AI Act supports the HLEG's guidelines by stating *"AI systems shall be developed and used in a way that allows appropriate traceability and explainability while making humans aware that they communicate or interact with an AI system as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights"*.

Transparency involves detailed documentation of an AI system's entire lifecycle along with the underlying operations that dictate its functioning. Ensuring transparency from the very design of an AI system is crucial to eliminate any uncertainty about its functionality and its application by those using it for clinical decisions. Not by chance, transparency is a core element of the FUTURE-AI guidelines [4, 5], being undertaken with the three guiding principles of Traceability, Explainability, and Usability. Transparency also ensures that an AI system is designed to be reproducible and auditable, laying the groundwork for accountability and responsibility. In the realm of academia, the push for transparency in AI aligns with well-known principles of *open data* and *open science* [6]. Yet, in the private sector, achieving transparency remains a complex issue, often due to competitive dynamics within the industry. In this brief discussion paper, we outline the key facets of transparency in medical imaging and provide a summary of the approaches being implemented in the EU H2020 ProCAncer-I project[2].

2. The implications of AI transparency in oncologic imaging

In the field of oncologic imaging, AI-based methodologies are increasingly dependent on data-driven approaches that manage large-scale, multimodal datasets efficiently. Specifically, in prostate cancer diagnostics, multiparametric magnetic resonance imaging (MRI) plays a critical role in detecting the presence of tumors and providing insights into tumor phenotypes [7, 8, 9]. However, for a comprehensive assessment of patient risk and condition, it is imperative to

[1] https://plato.stanford.edu/entries/trust/
[2] https://www.procancer-i.eu/

integrate these imaging data with clinical information, such as hormone levels, demographic details, and medical history. Additionally, the source of imaging data often varies, emanating from different clinical institutions with disparate clinical standards (e.g., PIRADS versions), acquisition protocols, and equipment from various manufacturers. Such variability has been shown to affect the efficacy of AI-driven tools, echoing the influence of heterogeneous population characteristics [10].

Given this intricate landscape, addressing transparency in AI for medical imaging demands a holistic approach. It requires diligent practices and robust technical measures to tracing and keep track of all the relevant choices across the entire AI life-cycle, from the initial data gathering phase through to the development, deployment, and operational stages of the system. Such an all-encompassing approach is vital for effectively managing the challenges posed by varied data origins, changing clinical guidelines, and the integration of imaging with clinical data for holistic analyses. In this context, the pursuit of transparency entails the establishment of a traceability system that serves as a definitive indicator of integrity and responsibility for both the AI technologies and the data on which they operate, thereby ensuring every phase is marked by clearness and accountability.

Transparency, in this scenario as in any other critical one, is articulated through the following key dimensions:

- *Data Traceability*: this involves meticulous documentation of the data origins, including details about who owns the data, how it was gathered, the clinical standards adhered to during collection, steps taken during data curation, locations and methods of data storage, and any pre-processing activities carried out. Such detailed record-keeping ensures every piece of data can be tracked through its life-cycle.
- *AI System Traceability*: this aspect covers the comprehensive and methodical documentation concerning the development, validation, and testing processes of the AI system. A standardized reporting format ensures that every step in the creation and application of the AI system can be audited and reviewed.
- *Decision Transparency*: this focuses on elucidating the AI system's decision-making processes. By providing clear explanations of the logic and rationale underpinning the AI's classifications or predictions, healthcare professionals are empowered to make informed decisions based on AI insights.

The traceability of data relies on well-established guidelines concerning data provenance and reuse. In contrast, traceability for AI models is still in the process of comprehensive development, despite some functionalities being partially integrated into MLOps frameworks. Decision transparency, a concept gaining momentum within the domain of eXplainable AI (XAI), represents a renewed emphasis on the decision-making process, drawing from a long-lasting pursuit [11, 12]. XAI is instrumental in facilitating productive collaboration between humans and AI systems. However, successful implementation necessitates the application of techniques from a broad array of fields, including human-computer interactions. Establishing standardized explanations from AI systems is paramount, ensuring they are robust and aligned with user needs to empower individuals and grant them complete control over the system.

3. Transparency measures in ProCAncer-I

The ProCAncer-I project aims to build the largest database of anonymized multiparametric MRI images related to prostate cancer, while adhering to the regulations outlined in the European Union General Data Protection Regulation (GDPR). The project's scope encompasses a spectrum of clinical scenarios, ranging from the diagnosis and characterization of prostate cancer to predicting responses to treatment and the likelihood of side effects post-treatment.

To this end, the clinical partners participating to the ProCAncer-I consortium have meticulously outlined the collection requirements for all clinical, imaging, pathology, and follow-up data. These requirements encompass essential clinical details that must accompany the images, such as prostate-specific antigen levels, biopsy outcomes, and confirmations of prostate cancer through prostatectomy reports. Additionally, specific guidelines have been established to capture vital information related to the medical images, aligned with the unique needs of each clinical scenario. This alignment ensures the development of an AI model that can effectively address the objectives set forth for the project.

Utilizing these multimodal data, the technical partners are currently focused on creating AI models capable of addressing the primary clinical tasks outlined in the project.

As part of this effort, a dedicated project activity is focused on ensuring trustworthiness and transparency by addressing all three dimensions mentioned above.

3.1. Data traceability

Within the project, medical and clinical data have undergone *FAIRification*, conforming to the principles of being Findable, Accessible, Interoperable, and Reusable. This process has been facilitated through a GDPR-compliant project infrastructure, utilizing the MOLGENIS metadata platform. This platform serves as the central metadata repository, enabling users to search for clinical and imaging metadata and assemble cohorts based on various variables. To represent the multimodal dataset, the Observational Medical Outcomes Partnership (OMOP)-Common Data Model (CDM) framework was expanded to include standardized imaging attributes [13]. This enhancement enabled streamlined cohort identification by leveraging DICOM metadata for the training and quality assurance of AI models. Furthermore, the extension encompassed refinements in the curation processes, establishing connections between the original and curated images through the utilization of standardized vocabularies.

3.2. AI System traceability

AI system traceability requires comprehensive documentation of the entire development process of an AI model or system, justifying and tracing it back to the data used for training and validation, the involved contributors, as well as the processing and refinement steps undertaken. This information is encapsulated in what we refer to as the *AI Model Passport*, which is housed within a designated model registry.

In drafting the content of the AI Model Passport, an exhaustive review of existing literature on AI traceability was conducted. The aim was to identify any available solutions, potential gaps, or shortcomings. The literature review categorized the identified works and approaches into three primary groups:

- Data and model provenance schemas
- Existing traceability tools
- Guidelines and recommendations

The initial focus was on provenance models designed to track the development of an AI model by archiving its history in connection with the data elements involved in its creation and the processes contributing to its evolution. Several models, particularly those centered on data, have emerged within the context of Open Science and data FAIRification. Notably, models such as DublinCore and CRMDig have already established themselves as standards. While various models addressing AI model lineage are available, widespread adoption and standardization remain limited. Notably, leading tech companies offer advanced models, though they have yet to be universally recognized as standards.

The investigation continued with an exploration of existing tools devised or conceptualized to facilitate AI traceability. Progress in this domain is rapid, with advancements stemming from areas beyond AI, such as software development and Continuous Integration and Continuous Deployment (CI/CD) life-cycles. Nonetheless, current tools fall short of providing all-encompassing end-to-end traceability support. Achieving this level of functionality would demand significant generalization capabilities and a comprehensive analysis of all pertinent factors.

The inquiry also extended to the latest traceability recommendations and guidelines. These guidelines delineate the essential information required for documenting the development, deployment, and usage of AI solutions. A cohesive proposal outlining the entities, components, and tags to be recorded is necessary to ensure comprehensive and meticulous record-keeping.

3.2.1. ProCAncer-I AI Model Passport

The structure, content, and organization of the ProCAncer-I AI Model Passport were defined through a comprehensive analysis of the literature as summarized above, along with an in-depth examination of data, AI/ML model provenance schemas, and the existing traceability tools.

A *divide-et-impera* strategy was adopted by scrutinizing each phase of the AI development and deployment chain individually, starting from data selection to model in-production monitoring. Subsequently, we worked to delineate the content required for the Passport for each of these stages.

The ProCAncer-I AI Model Passport has been formulated as a *minimal provenance model schema* comprising information aligning with the various stages of the AI chain:

1. The data collection process, tracking dataset characteristics and data localization.
2. The data processing pipeline, detailing data transformations from raw data to harmonized datasets.
3. The model training and validation process, including features extracted from data, training parameters, evaluation metrics, test conditions, and general AI/ML model characteristics for user monitoring and reproducibility.
4. The model operation and monitoring, storing performance metrics, uncertainty estimation, and metrics for detecting performance changes.

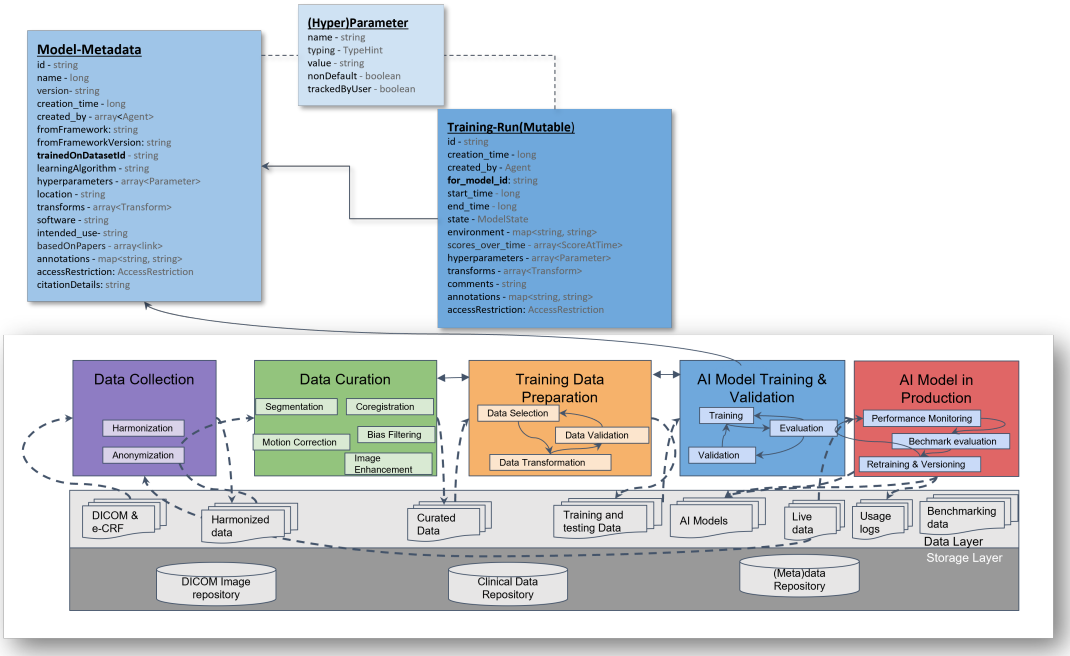

Figure 1: Phases of the AI development and deployment chain with a sample schema initially modeled for the AI Training phase.

For each phase, we analyzed existing schemas, their content, and ontologies to identify existing information. This guided the creation of an initial metadata schema with essential items required for each stage. The phased organization and an example of the initial schema are depicted in Figure 1.

Real examples of AI models and pipelines developed within the project were used to ensure the drafted set included all necessary items and considered all options in the ontologies and vocabularies. This comprehensive consideration involved iterative rounds of consultation between Passport designers and AI developers, documented through written specification documents. Figure 2 illustrates the initial version of the Passport featuring vision transformers developed as part of the project.

The current version of the passport has already been seamlessly integrated with industry-leading tools for data governance and management, such as DVC, and MLOps frameworks, like MLflow. This integration automates the population of metadata fields from these tools, reducing the need for manual input significantly. While the integration was an intensive effort, the results represent a groundbreaking solution.

3.3. Decision or algorithmic transparency

Research in Explainable Artificial Intelligence (XAI) aims to equip human decision-makers with insights into the operational logic of AI systems, particularly regarding their decision-making processes. However, the inherently data-driven nature of these algorithms means they

AI model Training	
Training framework	PyTorch
Training framework version	1.0
Training algorithm	Backpropagation on Stochastic Gradient Descent
Optimization algorithm	Adam (*/Rprop/ RAdam*)
Optimization algorithm parameters	Learning rate: 1e–4 Weight decay: none Number of epochs: 20 Loss function: Cross Entropy *(/Binary Cross Entropy/ Mean Squared Error/ Negative Log Likelihood…)* Batch size: 4 Dropout probability: none (*/none /value*)
Optimization algorithm parameter setting	Default (*/lr =….; /wdecay =….*)
Optimization algorithm parameter search	N (*Y/N*)
Optimization algorithm parameter search type	Random (*Grid/Random/Ad-hoc method*)
Train-Test process	k-fold cross validation with k=5 *(/k-CV with k value /none*)
Reproducibility flag	Y(*Y/N*)
Reproducibility parameter	42 (*none; value*)
Trained on DatasetUID	PCA-Train-692056885CBD1214B42368897C40A52F362F3F551EA6
Framework	
Framework version	
Libraries	
Architectural Hyperparameters Optimized	AI model type: Transformers (*/CNN /UNet /VisionTransformer /…*) Patch size: …, MLP size (d): …, hidden size (D): …, number of layers (L):…, number of attention heads:….
Architectural Hyperparameters optimization	Grid search (*Grid/random/ad-hoc method search*)

Figure 2: An example instantiating the ProCAncer-I AI Model Passport for a Vision Transformer developed in the project, with possible options for each item in brackets.

operate based on implicit problem specifications learned during training, which might not be immediately transparent or interpretable to humans. From the inception of the ProCAncer-I project, a concerted effort was made by both technical and clinical partners to delineate an optimal strategy for achieving explainability. This strategy required careful consideration of the various model types being developed (such as deep learning or radiomics) and the specific clinical tasks being addressed (for example, segmentation, detection, characterization). The process involved a comprehensive evaluation of the different use cases and AI methodologies to strike a balance that met the needs and expectations of both AI developers and clinical end-users.

Upon thorough analysis, a consensus was reached to integrate both local and global explainability techniques. This dual approach aims to assess and validate the AI model's performance from a broad perspective while providing targeted explanations for individual predictions or decisions. The exploration of XAI methods for both radiomics and DL models revealed specific challenges and necessitated tailored approaches for each. In this respect, a significant initiative undertaken by the clinical and technical partners within the consortium was the organization of a thematic session on AI explainability, conducted during a Consortium Plenary meeting. The purpose of this session was threefold: to gauge the clinical partners' understanding of explainability and the various XAI methodologies, to discern their expectations concerning the explainability and interpretability of AI models, and to solicit their preferences regarding explanation modalities. Concerning the last objective, when exploring preferences for explanation

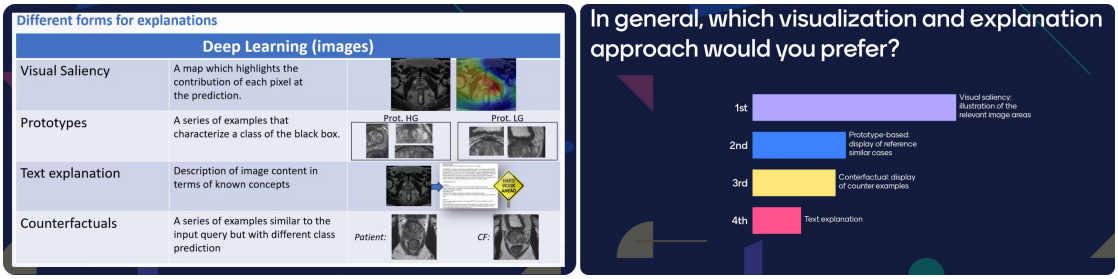

Figure 3: Left: types of explanations and visualization modalities for deep learning models. Right: Ranking of participants' preference for each option

modalities in the context of a deep learning models applied to general image-based prediction tasks (as shown in Figure 3 on the left), the participants demonstrated a clear preference for explanations that employed visual saliency maps to highlight areas of the image significantly impacting the model's prediction. This approach was the most favored, with prototype-based explanations ranking as the subsequent preferred option. These findings are going to be reflected in the implementation choices for the visualization and provisioning of explanations.

4. Conclusions

The ProCAncer-I project is dedicating significant resources to guarantee that the AI models crafted within the project garner trust and acceptance among the involved clinical partners but also across the broader clinical community and patients. In this respect, the introduction of the AI Model Passport emerges as an innovative strategy designed to streamline compliance with transparency directives, will be more and more stringent in the near future.

Acknowledgments

The work was partially supported by by the ProCancer-I European Union's H2020 program under Grant Agreement No. 952159 and the Tuscany Region project NAVIGATOR funded and supported by Bando Ricerca Salute Regione Toscana 2018 (DD 15397/2018).

References

[1] L. Marti-Bonmati, D.-M. Koh, K. Riklund, M. Bobowicz, Y. Roussakis, J. C. Vilanova, J. J. Fütterer, J. Rimola, P. Mallol, G. Ribas, et al., Considerations for artificial intelligence clinical impact in oncologic imaging: an ai4hi position paper, Insights into Imaging 13 (2022) 89.

[2] HLEG, High level expert group. ethics guidelines for trustworthy ai, https://tinyurl.com/4tej3t38, 2019. Accessed: 2024-03-15.

[3] EC, Artificial intelligence act, legislative resolution of 13 march 2024 on the proposal for a regulation of the european parliament and of the council on laying down harmonised

rules on artificial" p9 ta(2024)0138, https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf, 2024. Accessed: 2024-03-15.

[4] K. Lekadir, R. Osuala, C. Gallin, N. Lazrak, K. Kushibar, G. Tsakou, S. Aussó, L. C. Alberich, K. Marias, M. Tsiknakis, S. Colantonio, N. Papanikolaou, Z. Salahuddin, H. C. Woodruff, P. Lambin, L. Martí-Bonmatí, Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging, 2023. arXiv:2109.09658.

[5] K. Lekadir, A. Feragen, A. J. Fofanah, A. F. Frangi, A. Buyx, A. Emelie, A. Lara, A. R. Porras, A.-W. Chan, A. Navarro, et al., Future-ai: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare, arXiv preprint arXiv:2309.12325 (2023).

[6] R. Borgheresi, A. Barucci, S. Colantonio, G. Aghakhanyan, M. Assante, E. Bertelli, E. Carlini, R. Carpi, C. Caudai, D. Cavallero, et al., Navigator: an italian regional imaging biobank to promote precision medicine for oncologic patients, European radiology experimental 6 (2022) 53.

[7] E. Bertelli, L. Mercatelli, C. Marzi, E. Pachetti, M. Baccini, A. Barucci, S. Colantonio, L. Gherardini, L. Lattavo, M. A. Pascali, et al., Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric mri, Frontiers in oncology 11 (2022) 802964.

[8] E. Pachetti, S. Colantonio, M. A. Pascali, On the effectiveness of 3d vision transformers for the prediction of prostate cancer aggressiveness, in: International Conference on Image Analysis and Processing, Springer, 2022, pp. 317–328.

[9] E. Pachetti, S. Colantonio, 3d-vision-transformer stacking ensemble for assessing prostate cancer aggressiveness from t2w images, Bioengineering 10 (2023) 1015.

[10] N. M. Rodrigues, J. G. de Almeida, A. S. C. Verde, A. M. Gaivão, C. Bilreiro, I. Santiago, J. Ip, S. Belião, R. Moreno, C. Matos, et al., Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data, Computers in Biology and Medicine (2024) 108216.

[11] S. Colantonio, M. Martinelli, D. Moroni, O. Salvetti, D. Perticone, A. Sciacqua, F. Chiarugi, D. Conforti, A. Gualtieri, V. Lagani, Decision support and image & signal analysis in heart failure, Proc. of HEALTHINF. Madeira (2008) 288–295.

[12] F. Chiarugi, S. Colantonio, D. Emmanouilidou, M. Martinelli, D. Moroni, O. Salvetti, Decision support in heart failure through processing of electro-and echocardiograms, Artificial intelligence in medicine 50 (2010) 95–104.

[13] V. Kalokyri, H. Kondylakis, S. Sfakianakis, K. Nikiforaki, I. Karatzanis, S. Mazzetti, N. Tachos, D. Regge, D. I. Fotiadis, K. Marias, 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 7 (2023) e2300101.