Check for updates

CASE STUDY

REVISED How the First Medical Imaging Cancer Atlas EUCAIM Was Populated: The Experience of a Reference Hospital.

[version 3; peer review: 3 approved]

Ana Penadés Blasco[1], Leonor Cerdá Alberich[1], Ana de Marco García[1], Carina Soler Pons[1], Irene Marín Radoszynski[1], Ricard Martínez[2], Damián Segrelles-Quilis[3], Ignacio Blanquer[3], Luis Martí-Bonmatí[1,4]

[1]Biomedical Imaging Research Group (GIBI230), La Fe Health Research Institute, Valencia, Spain, 46026, Spain
[2]IRTIC- Instituto Universitario de Investigación de Robótica y Tecnologías de la Información y Comunicación, Valencia, Spain, 46980, Spain
[3]Instituto de Instrumentación para la Imagen Molecular, Universitat Politècnica de València, Valencia, Spain, 46011, Spain
[4]Medical Imaging Department, La Fe University and Polytechnic Hospital, Valencia, Spain, 46026, Spain

v3 **First published:** 10 Oct 2025, **5**:310
https://doi.org/10.12688/openreseurope.21016.1
Second version: 24 Nov 2025, **5**:310
https://doi.org/10.12688/openreseurope.21016.2
Latest published: 08 Dec 2025, **5**:310
https://doi.org/10.12688/openreseurope.21016.3

Abstract

The fragmentation and decentralization of medical data, including radiological imaging, continue to challenge large-scale observational research across Europe. Artificial Intelligence (AI) applied to big datasets is transforming diagnosis and treatments towards precision medicine across many diseases, yet the lack of findable, accessible, and interoperable datasets still limits model development, validation, and final clinical translation. The European Federation for Cancer Images (EUCAIM) project was launched in 2023 to address these challenges by establishing a secure centralized and federated infrastructure for the secondary use of large-scale oncological imaging and related clinical data.

By consolidating fragmented datasets, EUCAIM lays the groundwork for harmonized data governance and trusted cross-border sharing. Implementing a robust documentation framework is essential to ensure regulatory compliance, safeguard data integrity, and support secure data flows across institutional and national boundaries, fully aligned with European regulations and ethical standards.

EUCAIM builds on the AI for Health Imaging (AI4HI) initiative (Predictive In-silico Multiscale Analytics to support cancer personalized diagnosis and prognosis, empowered by imaging biomarkers -

Open Peer Review

Approval Status ✓ ✓ ✓

	1	2	3
version 3 (revision) 08 Dec 2025	✓ view		✓ view
version 2 (revision) 24 Nov 2025	✓ view		
version 1 10 Oct 2025	? view	✓ view	? view

1. **Deepak Kumar Shukla**, Rutgers University Newark Business School (Ringgold ID: 33869), Newark, USA
Neha Neha, Kent State University, Kent, USA

2. **Michal Strzelecki**, Lodz University of Technology, Lodz, Poland

3. **Ana M. Barragán Montero**, Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, UCLouvain, UCLouvain, Belgium

PRIMAGE, Accelerating the lab to market transition of AI tools for cancer management - CHAIMELEON, Novel pan-European imaging platform for artificial intelligence advances in oncology - EuCanImage, An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum - ProCancer-I, A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer - INCISIVE and integrates over 94 partners and more than 180 stakeholders spanning medical imaging, high performance computing, data standardization, innovation, and legal compliance. This large collaborative ecosystem reinforces EUCAIM's role as a reference for General Data Protection Regulation (GDPR) and European Health Data Space Regulation (EHDSR) adherence.

This publication presents the real-world experience of integrating imaging and clinical data from a reference university hospital into the EUCAIM infrastructure. It outlines the procedural, ethical, and legal challenges encountered, and details the strategies implemented to ensure compliance with data protection regulations, including privacy, security, and ethical standards. These insights offer a practical framework for future large-scale oncological imaging datasets harmonization and AI development, contributing to scalable, reproducible, and legally compliant research that strengthens Europe's capacity for trustworthy AI-driven oncology solutions.

Any reports and responses or comments on the article can be found at the end of the article.

Plain Language Summary

The EUCAIM (European Federation for Cancer Images) project was launched in 2023 to improve the way cancer imaging and related clinical data are shared and reused across Europe. Medical data, such as radiology images, are often fragmented and stored in different places, making it hard for researchers to access large, high-quality datasets. This limits the development, testing and validation of artificial intelligence (AI) tools that could help improve diagnosis and treatment.

EUCAIM aims to solve these difficulties by creating a secure and centralized infrastructure that brings together cancer imaging data from across Europe. It supports the safe and ethical reuse of these data for research, following strict rules on privacy, security, and informed consent. The project includes over 94 partners and 180 stakeholders from different fields (such as medical imaging, computing, data standards, and law) ensuring strong collaboration and extensive expertise.

EUCAIM builds on previous European projects (like PRIMAGE, CHAIMELEON, and ProCancer-I) and aligns with major EU regulations like the General Data Protection Regulation (GDPR) and the European Health Data Space Regulation (EHDSR). A key part of the project is ensuring that the infrastructure meets all legal and ethical

requirements for handling sensitive health data.

This article shares the real-life experience of a reference hospital contributing imaging and clinical data to EUCAIM and describes the technical and legal steps taken to ensure compliance and protect patients' information. These insights offer a practical example for other institutions considering participation in similar initiatives.

Overall, EUCAIM is helping to build a trusted European framework for using medical imaging data in AI research, aiming to accelerate innovation in cancer care while protecting patients' rights.

Keywords
federated infrastructures, sustainability, medical imaging, Artificial Intelligence, cancer research, data governance, innovation



This article is included in the Digital Health and Well-being collection.

Corresponding author: Ana Penadés Blasco (ana_penades@iislafe.es)

Author roles: Penadés Blasco A: Conceptualization, Data Curation, Investigation, Methodology, Project Administration, Writing – Original Draft Preparation; **Cerdá Alberich L**: Conceptualization, Investigation, Methodology, Validation, Writing – Review & Editing; **de Marco García A**: Conceptualization, Investigation, Methodology, Validation, Writing – Review & Editing; **Soler Pons C**: Data Curation, Methodology, Writing – Review & Editing; **Marín Radoszynski I**: Data Curation, Methodology, Writing – Review & Editing; **Martínez R**: Conceptualization, Investigation, Methodology, Validation, Writing – Review & Editing; **Segrelles-Quilis D**: Data Curation, Methodology, Writing – Review & Editing; **Blanquer I**: Data Curation, Methodology, Writing – Review & Editing; **Martí-Bonmatí L**: Conceptualization, Methodology, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project has received funding from the Digital Europe Programme (DIGITAL) under grant agreement No 101100633 (European Federation for Cancer Images [EUCAIM]).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Penadés Blasco A *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Penadés Blasco A, Cerdá Alberich L, de Marco García A *et al.* **How the First Medical Imaging Cancer Atlas EUCAIM Was Populated: The Experience of a Reference Hospital. [version 3; peer review: 3 approved]** Open Research Europe 2025, **5**:310 https://doi.org/10.12688/openreseurope.21016.3

First published: 10 Oct 2025, **5**:310 https://doi.org/10.12688/openreseurope.21016.1

REVISED **Amendments from Version 2**

This new revision introduces substantial improvements to strengthen contextualisation, clarify the technical framework, and enhance transparency regarding data governance and data usability within EUCAIM.

We have revised the Methods section to provide a clearer explanation of the two participation pathways available to institutions: Data Sharing Agreements (DSA) and Data Transfer Agreements (DTA). The revised text now details the rationale for offering both models, the institutional requirements for hosting federated nodes, the support provided by EUCAIM during installation and configuration, and the mechanisms that enable cross-site analytics when data remain local. The revision also clarifies the restrictions on raw data download, the role of Secure Processing Environments, and the technical safeguards that allow model training without data leaving the hospital.

Additionally, we have expanded the manuscript to clarify the documentation and tools referenced throughout the work, ensuring greater transparency and traceability.

Finally, the revision enhances the description of pseudonymisation versus anonymisation by explaining the two-layer privacy-preserving process, how re-identification risk is mitigated, and the safeguards implemented to prevent duplicate patient submissions across different projects.

Any further responses from the reviewers can be found at the end of the article

Introduction

Access to secondary use of medical imaging data for AI-driven research are hindered by fragmentation, lack of interoperability, and complex regulations across Europe regarding privacy and security issues. Although many research initiatives have created imaging repositories, these are often project-specific, temporally limited, and constrained in their potential for reuse. This is especially challenging in oncology[1–3], where AI tools for precise diagnosis, prognosis and treatment estimations require large, standardized and annotated datasets to reach full clinical applicability[4,5].

Multiple initiatives have previously generated valuable imaging repositories, examples include some Horizon Europe projects providing temporal available datasets such as Predictive In-silico Multiscale Analytics to support cancer personalized diagnosis and prognosis, empowered by imaging biomarkers - PRIMAGE, Accelerating the lab to market transition of AI tools for cancer management - CHAIMELEON, Novel pan-European imaging platform for artificial intelligence advances in oncology - EuCanImage, An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum - ProCancer-I, and A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer - INCISIVE. And also some existing repositories such as The Cancer Imaging Archive (TCIA), which provides curated datasets but largely centred on disease-specific collections; GrandChallenges, where datasets are created to support competitions and are rarely updated or expanded once each challenge is completed; and domain-focused initiatives such as DESIRE in radiotherapy, which target specific clinical use cases and remain confined to their original scope. These projects have demonstrated the scientific value of shared imaging data, but they lack the harmonised governance, long-term sustainability mechanisms, and structured clinical context that are required for large-scale, reproducible AI research. EUCAIM complements these existing resources since it provides a hybrid federated–centralised infrastructure with common standards for anonymisation, metadata, clinical linkage and data quality, enabling continuous expansion of the atlas and supporting cross-border and GDPR compliant analysis at scale

The European Federation for Cancer Images (EUCAIM) project addresses these challenges by developing a large-scale, secure, and federated digital infrastructure for medical imaging and related clinical data. EUCAIM accelerates AI research in personalized medicine, offering tools that support clinical decision-making in complex settings[6]. EUCAIM builds on the AI for Health Imaging (AI4HI) network, which includes five major Horizon Europe projects (PRIMAGE, CHAIMELEON, EuCanImage, ProCancer-I, and INCISIVE). With over 90 partners and more than 180 stakeholders, EUCAIM is creating a GDPR-compliant, sustainable ecosystem for AI-based cancer research[7,8].

While EUCAIM offers a unique opportunity to harmonize and share annotated oncological imaging data, its success depends on overcoming technical barriers (standardization, transformation, anonymization, curation), strict compliance with the General Data Protection Regulation (GDPR), the European Health Data Space Regulation (EHDSR), and the Artificial Intelligence Act (AI Act), and with a clear governance frameworks[9]. The EUCAIM platform aims to host over 60 million images from more than 100,000 patients, transforming medicaldata use in personalized actionable care[10].

EUCAIM's hybrid model allows Data Providers to transfer data to a reference node or set up their own federated node. This paper analyses the use case of real-world data transfer from a referral hospital, detailing workflows and practical solutions to secure legal compliant integration into EU-wide repositories.

Material and methods

Data integration in EUCAIM follows either a *Data Sharing Agreement* (DSA), where data stays within the institution via a federated node; or a *Data Transfer Agreement* (DTA), where data is physically moved to a reference node which comprises 10 Gigabyte R283-ZF0-AAL1 nodes, each equipped with two AMD EPYC 9474F 48-core processors (a total of 960 cores), 7.68 TB of RAM, 15 NVIDIA A30 and 10 NVIDIA L40S GPU accelerators with 48 GB of RAM each, as well as an additional storage server with 16 TB of NVMe SSD disks connected to the nodes via dual 25 GbE links. In the experiments performed in previous works[11], this capacity is sufficient to deal with a workload of over 30 concurrent users.

For the retrospective imaging data included in this study, no informed consent was required, an exemption of informed consent was submitted to and approved by the Ethics Committee. All these nodes are core to EUCAIM's infrastructure.

The choice between a DSA and a DTA reflects the heterogeneous technical capabilities and governance preferences of participating hospitals. Some institutions require that data processing remains fully under their control; for these cases, EUCAIM enables the deployment of a federated node within the hospital infrastructure, hosted on local servers and integrated with the EUCAIM platform through secure, encrypted channels. This environment operates as a Secure Processing Environment under the EHDS framework, allowing authorised users to run analytics and train models locally without any image data leaving the institution. EUCAIM provides technical specifications, containerised services, and remote support to assist hospitals in the installation and configuration of these nodes, although the hardware is typically procured and maintained by the institution according to its internal policies. Other hospitals opt for a DTA, transferring data to the EUCAIM reference node where harmonisation, storage and computation are centrally managed. While models trained within this environment may be exported subject to Access Committee approval, raw imaging data are never downloadable to external servers. This dual approach offers flexibility for centres with different resources and legal constraints while ensuring that all data, whether local or centralised, can be incorporated into cross-site analyses through a unified and privacy-preserving architecture.

Data extraction and exposure pipeline

To ensure standardized, de-identified, and compliant data integration into the EUCAIM infrastructure, a structured Extraction, Transformation, and Loading (ETL) pipeline was defined, incorporating best practices in data governance, security, and regulatory adherence. Imaging data were extracted from the hospital Picture Archiving and Communication System (PACS) through specific nodes in Digital Imaging and Communication in Medicine (DICOM) format according to each project's inclusion criteria. Corresponding clinical data were retrieved from the Electronic Health Record (EHR) system to maintain proper alignment.

The data preparation workflow includes a two-stage privacy-preserving process:

1. Local pseudonymisation:
Pseudonymisation is performed on a restricted-access Virtual Machine by authorized personnel of the Experimental Radiology and Imaging Biomarkers Platform (PREBI) using an in-house tool developed by the Biomedical Imaging Research Group (GIBI230). The process includes: replacement of PatientID, PatientName and AccessionNumber with specific pseudonym and hashes using Blake2b, renaming of folder structures to remove personally identifiable information (PII), removal of public and private DICOM metadata that may contain PII, exclusion of screenshots (ImageType = SCREEN SAVE) and manual review of secondary captures (DERIVED or SECONDARY). This step ensures consistent pseudonymised identifiers for repeated submissions of the same patient at the hospital level. The mapping between original and pseudonymised identifiers is maintained only temporarily by the hospital IT service and is not accessible externally.

2. EUCAIM anonymization:
Once pseudonymised data are transferred to the Pseudonymised Medical Imaging Repository, the EUCAIM anonymization pipeline is applied. This pipeline enforces the EUCAIM DICOM Anonymization Profile and generates a unique hash per patient, linked to the project and site, with no retained traceability. Sensitive DICOM headers are removed, and OCR-based pixel-level text detection can be applied to eliminate "burned-in" personal information. File integrity checks and pixel-level duplicate detection can be performed to prevent repeated inclusion of the same images across different projects.

This two-layer anonymization approach mitigates re-identification risks, aligning with procedures recommended by the Spanish Data Protection Agency and best practices from the Singaporean authority (Figure 1).

Additionally, after transferring it to Pseudonymised Medical Imaging Repository and anonymising the datasets, the DICOM File Integrity Checker developed by our group was used to detect corrupted or missing files. Data were finally ingested into EUCAIM by transferring them to the Reference Node through the QP-Insights API, or by sharing them through a federated node. For data standardization, all DICOM files and metadata were validated for compliance with EUCAIM's structure and interoperability standards.

The data preparation process and the tools involved are defined in the EUCAIM Handbook. Additionally, the metadata of all datasets were registered in the EUCAIM Public Catalogue,

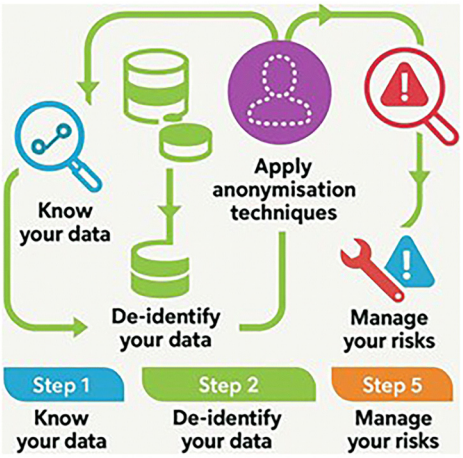

Figure 1. Steps for anonymization. Source: Adapted from AEPD. Guide to basic anonymization. Prepared by the National Data Protection Authority of Singapore (PDPC - Personal Data Protection Commission Singapore.

which follows the Health DCAT-AP standard and ensures compliance with FAIR principles at the dataset level. An additional layer of dataset discoverability is provided to EUCAIM Data Users through the Federated Query tool, which enables them to perform queries based on specific criteria and retrieve the number of cases (as aggregated numerical results) that meet those criteria

The EUCAIM Federated Node provides an open-source, fully integrated Data Lake, Registry, and Secure Virtual Research Environment, backed by dedicated computing resources. Its API enables the ingestion while logging all transfers for traceability and auditability. Federated nodes are fully integrated within the EUCAIM platform and meet the EHDSR requirements for Secure Processing Environments (SPEs). SPEs ensure that data processing stays local, under the health data holder's continuous control, significantly reducing transfer risks. Specialized mediation software enforces strict anonymization rules, restricts user actions, and logs all activities to prevent misuse and ensure full accountability.

By leveraging SPEs and federated nodes, EUCAIM provides a robust, compliant, and auditable framework for the secure secondary use of health data. This rigorous ETL pipeline ensures scalability, data quality through the access to tools and workflows within EUCAIM, and legal compliance, positioning EUCAIM's reference and federated nodes as the backbone for integrating data from research projects, clinical trials[12], and future initiatives, in line with European regulations such as the Data Act and Data Governance Act. This strategy guarantees EUCAIM's long-term sustainability and impact on AI research in medicine[13].

Documentation framework for data integration to EUCAIM

Ensuring compliant data integration within EUCAIM requires strict adherence to clear documentation protocols. At its current stage, decisions on the final integration of a dataset and/or federated node are verified by an Access Committee (AC). During this phase, it is mandatory to provide documentary evidence that guarantees accountability in EUCAIM's operation and demonstrates the application reliability.

For data transfers to EUCAIM, the following documents are considered to safeguard data integrity and regulatory compliance previous to the signature of the DTA:

- Data Protection Officer (DPO) Report or Self-Declaration: formal declaration by the transferring institution affirming that the data processing activities comply with GDPR requirements, including data minimization, data protection by design, and risk mitigation measures.
- Codes of conduct or certification schemes adherence: a non-mandatory certification of GDPR declaring the institution provides evidence of adherence to established codes of conduct or certifications relevant to GDPR compliance.
- Data Protection Impact Assessment (DPIA). documented evidence of the completed DPIA must be submitted in cases where, in accordance with Article 25 of the GDPR and the relevant national authority's whitelist, it is required.
- Ethical Approval Documentation from the relevant Ethics Committee: certifying that the data transfer aligns with ethical research practices and data protection regulations. The requirement for ethics approval, its specific content, or any granted exemption is determined by national law. In cases where ethics approval is not required, it is recommended to be explicitly documented in the statement provided by the DPO.
- Legal Representation Declaration: signed by the institution's legal representative, confirming that the data transfer adheres to legal obligations and that the representative is duly authorized to sign agreements on behalf of the institution. This documentary requirement is crucial, as requests might be handled by researchers or individuals who lack the legal capacity to make binding declarations of intent.

For Data Sharing Agreements (DSAs) via federated nodes, the same requirements apply, with an additional Security Report or certification (e.g., ISO 27001) document verifying that data sharing meets EUCAIM's security standards for encryption, user authentication, and access control.

All documentation must be verifiable and may undergo audits. Federated nodes must show secure operations and technical interoperability with the EUCAIM platform. Likewise, datasets claimed to be anonymized will be checked: any re-identification risk means the data will be returned for further processing until fully compliant. This framework guarantees that all partners contribute data responsibly, supporting secure, transparent, and efficient data sharing within the EHDSR.

The relevant legal documentation can be summarized in Figure 2

Use case of data transfer from completed research projects

The PRIMAGE project (GA: 826494) demonstrated how historical datasets could be successfully transferred and validated within the EUCAIM reference node. The main steps included:

- Legal and Ethical Compliance: all approvals and documentation were secured.
- Data Processing & Transfer: imaging data were extracted, anonymized, and reformatted into EUCAIM-compliant DICOM structures. Clinical data was similarly processed, aligned with the EUCAIM Construction, Design and Management (CDM) and hyper-ontology based on mCODE specifications.

Use case of data transfer from oncology clinical trials

The previous approach was also applied to ongoing clinical trials, expanding EUCAIM's reference node cases. Additional challenges emerged, particularly regarding data ownership and sponsor agreements. Key strategies included:

- Data Ownership Clarification: a legal framework defined that imaging data belong to the patient's EHR, not the trial sponsor, minimizing transfer conflicts and ensuring EU compliance.

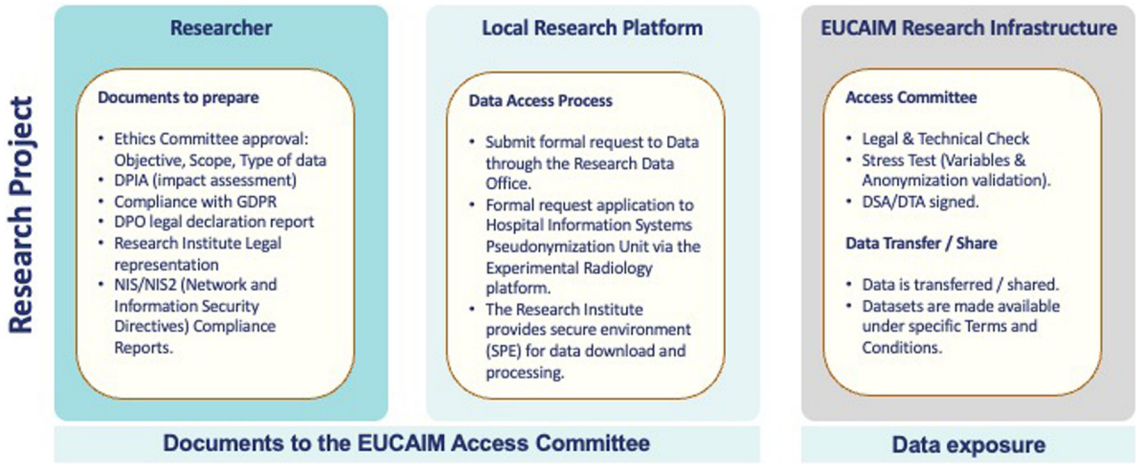

Figure 2. Summary of the relevant legal documentation in EUCAIM.

- Informed Consent: exemptions were requested as data were fully anonymized and restricted to research use.

- Periodic Scheduling: semi-annual transfers allow continuous integration in line with regulatory approvals.

Use case of data sharing from ongoing research projects

The CHAIMELEON project (GA: 952172) illustrated a continuous dataset sharing process without disrupting ongoing research, serving as a model for federated nodes. For these projects, a more dynamic workflow was required due to evolving data collection and regulatory requirements:

- Ethics Committee Approval: an addendum updated the original approval to permit secondary data use in EUCAIM.

- Data Harmonization & Monitoring: a real-time tracking system ensured seamless, compliant integration.

CHAIMELEON reached a high maturity level by preparing a robust DSA, anonymizing data, implementing a secure node, and completing a DPIA, setting a strong precedent for future compliant transfers.

Ethical and legal considerations

Each use case prioritizes full compliance with data protection regulations, especially GDPR. All data transferring and sharing activities were approved by the relevant ethics committees, ensuring confidentiality, integrity, and security. This rigorous approach has enabled the secure transfer and sharing of large volume of anonymized data, contributing to advanced AI research. Within EUCAIM, data use is strictly limited to users with approved research projects that meet institutional ethics standards and follow the Access Committee's protocols.

The recent approval of the EHDSR marks a significant milestone for EUCAIM and similar initiatives, providing a harmonized framework for secure access, sharing, and secondary use of large-scale medical datasets. This framework sets clear rules for research and innovation, supported by strict requirements for security, privacy, and informed consent, fostering an interoperable and sustainable data ecosystem. EUCAIM's activities have been progressively aligned with this evolving landscape, positioning as a pilot platform under HealthData@EU. Its formal integration into the EHDS ecosystem will strengthen data accessibility and protection, reinforcing the project's long-term sustainability and scientific value, as data volume and diversity grow.

EUCAIM addresses sustainability through the establishment of a European Digital Infrastructure (EDIC) to enable multi-country investments in large-scale projects. The proposed model combines national and node contributions with European funding to ensure long-term operation of the Central Hub and integration of new partners.

Results

The implementation of a dedicated, harmonized pipeline for data transfer and sharing enabled the successful ingestion of cancer imaging data into the EUCAIM infrastructure. This process established a replicable and scalable workflow suitable for future large-scale federated data integration.

In total, 12,484 medical imaging studies at our hospital and research institute were identified, reviewed, and prepared for integration into EUCAIM, covering both research datasets and real-world clinical trial data. Of these, 10,892 studies (87.24%) from 6,105 patients have been fully processed and validated for upload through either a reference or federated node approach. Our institution has contributed to several initiatives, including the completed PRIMAGE project with 878 studies,

the ongoing CHAIMELEON project with 5,346 studies, and a series of Oncology Clinical Trials comprising 4,668 studies. The datasets include multiple modalities ((Computed Tomography - CT, Magnetic Resonance - MR, mammography, Positron Emission Tomography-Computed Tomography - PET-CT)) and represent a broad spectrum of cancer cases and patientpopulations (Table 1).

The ETL pipeline achieved a processing efficiency of 98.6%, with minimal data loss. Automated DICOM anonymization, combined with manual checks, ensured GDPR compliance while maintaining clinical relevance. PRIMAGE delivered one of EUCAIM's largest structured paediatric oncology datasets, setting a benchmark for rare disease integration. CHAIMELEON validated a stepwise model for harmonized data sharing, and the inclusion of ongoing clinical trial data demonstrated the feasibility of incorporating prospective datasets into the federated infrastructure.

Several challenges emerged that required refinements to optimize efficiency, scalability, and compliance. A key technical challenge was achieving interoperability across diverse imaging formats and legacy PACS metadata, which often required extensive transformation and custom mapping. Variability in clinical data mappings also highlighted the need for more automated and standardized processes.

Administrative and legal aspects added complexity. Ethics approvals often needed significant revisions depending on the project type, and requests for informed consent exemptions required careful justification and additional documentation. These barriers underscored the lesson learned: the EHDS will require organisations to renew their ethical and legal governance models to ensure predictable, efficient data sharing, including data holders, research infrastructures, and users.

Scalability was initially hindered by reliance on manual anonymization checks, slowing processing and introducing inconsistencies. Introducing automated validation tools and batch workflows significantly improved efficiency, though manual review remains necessary for nuanced cases like secondary capture images.

These insights guided refinements to our protocols. Automated metadata validation minimized errors, while standardized ethics templates helped accelerate approval timelines. Experience gained across projects emphasized the importance of early alignment with EUCAIM requirements, ensuring technical, compliance, and governance aspects are clear from the outset.

Despite initial hurdles, this process strengthened our capacity to securely transfer oncological imaging data with the highest standards of security, privacy, and interoperability, providing a solid foundation for future data-sharing efforts within EUCAIM and similar large-scale initiatives.

Discussion

EUCAIM is established as a landmark initiative in oncological imaging, offering Europe's first large-scale hybrid infrastructure for the secondary use of imaging, clinical, and molecular data. By integrating datasets from completed research, ongoing projects, and clinical trials, EUCAIM ensures alignment with the EHDSR and GDPR, thereby supporting the full potential of artificial intelligence in advancing predictive and personalized cancer care. Nevertheless, significant challenges remain, particularly data fragmentation, legacy system integration, and cross-border regulatory complexity.

A key barrier to building a unified repository has been the variability in data formats, metadata, and institutional governance practices. EUCAIM addresses this challenge through structured documentation frameworks and standardized ETL pipelines, which enhance data governance and promote interoperability across systems. Even so, harmonizing legacy PACS data and oncological metadata will continue to demand refinements and shared best practices.

From the outset, ethical and legal considerations have been central. Multi-step anonymization, robust audit logs, and clear governance models ensure compliance and traceability. However, achieving consistent ethics approvals and seamless cross-border data sharing still requires effort. The EHDS framework will be instrumental in easing these processes, but only if health data holders, research infrastructures, and users renew and align their governance models.

Table 1. Overview of prepared Data Sources for Integration into the EUCAIM Reference Node and Federated Node.

Acronym	Project status	Number of identified imaging studies	Number of prepared imaging studies	Node
PRIMAGE	Completed	878	878	Reference
CHAIMELEON	Ongoing	5,346	5,346	Federated
Clinical Trials	Completed and ongoing	6,260	4,668	Reference

The lessons learned from large-scale data integration reinforce the value of early legal and technical alignment. PRIMAGE, CHAIMELEON and ongoing clinical trials each illustrate how proactive regulatory planning and clear data management frameworks enable scalable, compliant data sharing.

Long-term sustainability will rest on strict adherence to Findable, Accessible, Interoperable and Reusable (FAIR) principles, supported by reference and federated nodes, FAIR Data APIs, and persistent identifiers that guarantee data findability and reusability. Automating compliance checks and deploying secure cloud-based environments will be essential to handle ever-growing data volumes, while standardized metadata schemas will further support seamless reuse.

Through the continuously refinement of its governance strategies and proactive alignment with the evolving regulatory landscape under the EHDS, EUCAIM is positioned to become a cornerstone for AI-driven oncology research. Its robust infrastructure not only accelerates the clinical translation of AI tools but also strengthens a culture of trustworthy, collaborative data sharing across Europe, reinforcing EUCAIM's pivotal role in shaping the future of cancer imaging and precision medicine. Central to this effort is the essential contribution of data providers, whose engagement ensures the availability, diversity, and quality of datasets reinforcing EUCAIM in shaping the future of precision medicine.

Ethical approval statement

This manuscript has been approved by the Research Ethics Committee on Medicinal Products of La Fe University and Polytechnic Hospital under registration number 2022-437-1. For the retrospective imaging data included in this study, no informed consent was required, an exemption of informed consent was submitted to and approved by the Ethics Committee.

Data availability

No data are associated to this manuscript.

Acknowledgments section

Patricia Serrano Candelas, Silvia Flor Arnal, Lucas Espuig Peiró, Antonio Orduña Galán, Cayetano Hernandez Marín and Javier Medina Álvarez.

References

1. Genovese S, Bengoa R, Bowis J, *et al.*: **The European Health Data Space: a step towards digital and integrated care systems.** *J Integr Care.* 2022; **30**(4): 363–372.
Publisher Full Text
2. Lehne M, Sass J, Essenwanger A, *et al.*: **Why digital medicine depends on interoperability.** *NPJ Digit Med.* 2019; **2**(1): 79.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
3. Pierce HH, Dev A, Statham E, *et al.*: **Credit data generators for data reuse.** *Nature.* 2019; **570**(7759): 30–32.
PubMed Abstract | **Publisher Full Text**
4. Norgeot B, Glicksberg BS, Butte AJ: **A call for deep-learning healthcare.** *Nat Med.* 2019; **25**(1): 14–15.
PubMed Abstract | **Publisher Full Text**
5. Koh DM, Papanikolaou N, Bick U, *et al.*: **Artificial Intelligence and machine learning in cancer imaging.** *Commun Med (Lond).* 2022; **2**(1): 133.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
6. Radclyffe C, Ribeiro M, Wortham RH: **The assessment list for trustworthy Artificial Intelligence: a review and recommendations.** *Front Artif Intell.* 2023; **6**: 1020592.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
7. Gagliardi D, *et al.*: **AI4HI: AI and health imaging research priorities.** *Eur J Radiol Open..* 2023.
8. Kondylakis H, Ciarrocchi E, Cerda-Alberich L, *et al.*: **Position of the AI for Health Imaging (AI4HI) network on metadata models for imaging biobanks.** *Eur Radiol Exp.* 2022; **6**(1): 29.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
9. Voigt P, Von der Bussche A: **The EU General Data Protection Regulation (GDPR)**. A Practical Guide, 1st Ed., Cham: Springer International Publishing, 2017; **10**(3152676): 10–5555.
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR guiding principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**(1): 160018.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
11. Segrelles Quilis JD, Lozano P, Blanco-Sanchez A, *et al.*: **Experiences on using the CHAIMELEON secure processing environment in an open competition addressing five AI challenges**.
Publisher Full Text
12. Gyrard A, Abedian S, Gribbon P, *et al.*: **Lessons learned from European health data projects with cancer use cases: implementation of health standards and internet of things semantic interoperability.** *J Med Internet Res.* 2025; **27**: e66273.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**
13. Martí-Bonmatí L, Blanquer I, Tsiknakis M, *et al.*: **Empowering cancer research in Europe: the EUCAIM cancer imaging infrastructure.** *Insights Imaging.* 2025; **16**(1): 47.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 3

Reviewer Report 01 January 2026

https://doi.org/10.21956/openreseurope.23947.r65769

© **2026 Montero A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Ana M. Barragán Montero**
Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, UCLouvain, UCLouvain, Belgium

The authors have answered all my comments, very good job! Congratulations on this great initiative.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial intelligence, radiation therapy, medical imaging, cancer treatment

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 30 December 2025

https://doi.org/10.21956/openreseurope.23947.r65770

© **2025 Neha N et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Deepak Kumar Shukla**
Rutgers Business School, Rutgers University Newark Business School (Ringgold ID: 33869), Newark, Jersey, USA

Neha Neha
Kent State University, Kent, USA

Author has made all the suggested changes.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial IntelligenceComputer VisionImage ProcessingDeep learningMedical Imaging

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 02 December 2025

https://doi.org/10.21956/openreseurope.23735.r64762

© **2025 Neha N et al.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Neha Neha
Kent State University, Kent, USA
Deepak Kumar Shukla
Rutgers Business School, Rutgers University Newark Business School (Ringgold ID: 33869), Newark, Jersey, USA

Author has made all the suggested updates.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial IntelligenceComputer VisionImage ProcessingDeep learningMedical Imaging

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 22 November 2025

https://doi.org/10.21956/openreseurope.22734.r62149

© **2025 Montero A.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Ana M. Barragán Montero**
Molecular Imaging, Radiation and Oncology (MIRO) Laboratory, UCLouvain, UCLouvain, Belgium

Review of the manuscript Open Research Europe 2025, 5:310 "How the first medical imaging cancer atlas EUCAIM was populated: the experience of a reference hospital"
The manuscript describes the EUCAIM ecosystem and the methodology to submit images, illustrated with the experience of a hospital in the consortium. The article is well written and clear, and it is suitable for publication in the journal. However, I have some comments that I believe can help to increase the quality of the manuscript and before final publication.
MAJOR COMMENTS

1. **Improve the contextualisation and illustrate the state of the art. In the first paragraph, they describe the current ecosystem for data sharing and the limitations, highlighting the need for an initiative like EUCAIM. While the paragraph is well written, I am missing a few concrete examples and references that support the arguments given in the paragraph. For instance, the authors mention that "many research initiatives have created imaging repositories, but they are often project-specific, temporally limited, …" but no example is given. Please provide at least two or three examples. Also what is the potential of EUCAIM with respect to other existing initiatives for large-scale data collection like TCIA (https://www.cancerimagingarchive.net/), DESIRE (https://www.straaleterapi.dk/en/desire/about-desire/), GrandChallenges (https://grand-challenge.org/challenges/), etc I think a paragraph discussing this and giving specific examples will help the reader to better see the potential and value of EUCAIM**

1. **Describe more in detail the system for hosting and using the data. The authors distinguish between two options (materials and method, paragraph 1): 1) DSA, where the data stays in the system, or DTA, where the data goes to the federated node. Can you provide more technical information about this part? Or either reference to already published protocols/papers describing this part? Some questions here: Why these two options and not only DTA directly? For hospitals choosing 1), is it up to them to buy a server to host the data? How is the connection then to other hospitals willing to use the data through the DSA? Does EUCAIM provide a technical team to make all these installations in the local computers or the hospital should do it? How can other hospitals use the DSA data to train models if the data did not leave the hospital? For DTA data, can I download the data into my servers to perform research (e.g. trained AI models)?**
   I understand that the paper here focuses on the population of the atlas, but some clarification (a paragraph will be enough) here is important, so that the community understands how the data can be used later. Otherwise, it seems a very nice platform to store data but the article does not give any clue about the potential to use the data later.

1. **Disclose or add reference for already available material. This is published in Open Research journal, so I guess that the goal is to make as open as possible the results presented here. There are many parts where you mention that the authors (or the EUCAIM team) has developed a lot of material, but no reference is given. For instance,**

in material and methods "For data standardization, all DICOM files and metadata were validated for compliance with EUCAIM structure", "The DICOM File Integrity Checker developed by our group" → is this material published elsewhere or made open-source through a github/github or similar repository? The amount of work that has been developed here has an incredible value for the community and should be made public whenever possible.

1. **Pseudonymisation versus Anonymisation. I have a question regarding this, since both are mentioned during the text, it is not clear to me if the data uploaded is pseudonymised, and there is a log file somewhere (in the hospital submitting the data) that keeps a link (e.g. PatientID) between the data submitted and the real patient identity, or the data is fully anonymised. If the data is totally anonymised and there is no log-file, how do we prevent errors in the future if the hospital submit the same patient (e.g. submission in 2025 for project X, new patient ID projectX_01 - submission in 2026 of same patient for project Y, new patient ID project Y_01). Also, besides removing DICOM tags, can you describe if any action on the images has been done or not? There are some people that advocate to remove facial information from head-and-neck scans, but this can affect the potential use of the data (e.g. in head-and-neck cancer patients).**

1. **Data quality and documentation. You do mention DICOM compliance, anonymisation, etc, but what about data quality and documentation?**
    1. **About data quality, is there any metadata or any quality check to measure the quality of the submitted data? I am thinking more on the annotations of the data, for instance, if the annotations have been done by experts and following consensus international guidelines, etc A concrete example would be the adherence to contouring guidelines in radiotherapy images. Data annotations not being compliant to these guidelines and used to train models might carry future problems. Also, we could think about data quality of the images, if a hospital submit images with very low resolution, artefacts, etc. A way to go would be to enable a community rating for each dataset ...**
    2. **About data documentation. I am putting myself on the user side, and the first question that comes to my mind is: how do I find and decide which data I use for my project? Having data documentation is very important, like a Data Sheet (similar to what they have done in https://arxiv.org/abs/1803.09010, https://datanutrition.org/, ...). Is there any similar documentation standard in EUCAIM?**

2. **In general some paragraphs feel too much "GPT-style", sometimes repeating the same but with different words. Please review and try to make the text less redundant and more concise whenever possible.**

Is the background of the case's history and progression described in sufficient detail?
Yes

Is the work clearly and accurately presented and does it cite the current literature?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Not applicable

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial Intelligence for medical imaging, radiotherapy treatment planning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 03 Dec 2025

Ana Penades Blasco

Response to the REVIEWER #3 COMMENTS
Thank you very much for your comments and feedback. We have worked on improving the manuscript with your suggestions and comments.
Reviewer #3:
Reviewer 3.1: Improve the contextualisation and illustrate the state of the art. In the first paragraph, they describe the current ecosystem for data sharing and the limitations, highlighting the need for an initiative like EUCAIM. While the paragraph is well written, I am missing a few concrete examples and references that support the arguments given in the paragraph. For instance, the authors mention that “many research initiatives have created imaging repositories, but they are often project-specific, temporally limited, …” but no example is given. Please provide at least two or three examples. Also what is the potential of EUCAIM with respect to other existing initiatives for large-scale data collection like TCIA (https://www.cancerimagingarchive.net/), DESIRE (https://www.straaleterapi.dk/en/desire/about-desire/), GrandChallenges (https://grand-challenge.org/challenges/), etc I think a paragraph discussing this and giving specific examples will help the reader to better see the potential and value of EUCAIM
Authors: Following the comments of the reviewer we now incorporated several sentences at the Introduction: “Multiple initiatives have previously generated valuable imaging repositories, examples include some Horizon Europe projects providing temporal available datasets such as Predictive In-silico Multiscale Analytics to support cancer personalized diagnosis and prognosis, empowered by imaging biomarkers - PRIMAGE, Accelerating the lab to market transition of AI tools for cancer management - CHAIMELEON, Novel pan-European imaging platform for artificial intelligence advances in oncology - EuCanImage, An

AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum - ProCancer-I, and A multimodal AI-based toolbox and an interoperable health imaging repository for the empowerment of imaging analysis related to the diagnosis, prediction and follow-up of cancer - INCISIVE. And also some existing repositories such as The Cancer Imaging Archive (TCIA), which provides curated datasets but largely centred on disease-specific collections; GrandChallenges, where datasets are created to support competitions and are rarely updated or expanded once each challenge is completed; and domain-focused initiatives such as DESIRE in radiotherapy, which target specific clinical use cases and remain confined to their original scope. These projects have demonstrated the scientific value of shared imaging data, but they lack the harmonised governance, long-term sustainability mechanisms, and structured clinical context that are required for large-scale, reproducible AI research. EUCAIM complements these existing resources since it provides a hybrid federated–centralised infrastructure with common standards for anonymisation, metadata, clinical linkage and data quality, enabling continuous expansion of the atlas and supporting cross-border and GDPR compliant analysis at scale."

Reviewer 3.2: Describe more in detail the system for hosting and using the data. The authors distinguish between two options (materials and method, paragraph 1): 1) DSA, where the data stays in the system, or DTA, where the data goes to the federated node. Can you provide more technical information about this part? Or reference to already published protocols/papers describing this part? Some questions here: Why these two options and not only DTA directly? For hospitals choosing 1), is it up to them to buy a server to host the data? How is the connection then to other hospitals willing to use the data through the DSA? Does EUCAIM provide a technical team to make all these installations in the local computers, or the hospital should do it? How can other hospitals use the DSA data to train models if the data did not leave the hospital? For DTA data, can I download the data into my servers to perform research (e.g. trained AI models)?
I understand that the paper here focuses on the population of the atlas, but some clarification (a paragraph will be enough) here is important, so that the community understands how the data can be used later. Otherwise, it seems a very nice platform to store data, but the article does not give any clue about the potential to use the data later**.**
Authors: To address this comment, we have included the following paragraph on page 6: "The choice between a DSA and a DTA reflects the heterogeneous technical capabilities and data governance preferences of participating hospitals. Some institutions require that data processing remains fully under their control; for these cases, EUCAIM enables the deployment of a federated node within the hospital infrastructure, hosted on local servers and integrated with the EUCAIM platform through secure, encrypted channels. This environment operates as a Secure Processing Environment under the EHDS framework, allowing authorised users to run analytics and train models locally without any image data leaving the institution. EUCAIM provides technical specifications, containerised services, and remote support to assist hospitals in the installation and configuration of these nodes, although the hardware is typically procured and maintained by the institution according to its internal policies. Other hospitals opt for a DTA, transferring data to the EUCAIM reference node where harmonisation, storage and computation are centrally managed. While software models trained within this environment may be exported subject to the Access Committee approval, original imaging data are never downloadable to external

servers. This dual approach offers flexibility for centres with different resources and legal constraints while ensuring that all data, whether local or centralised, can be incorporated into cross-site analyses through a unified and privacy-preserving architecture."

Reviewer 3.3: Disclose or add reference for already available material. This is published in Open Research journal, so I guess that the goal is to make as open as possible the results presented here. There are many parts where you mention that the authors (or the EUCAIM team) have developed a lot of material, but no reference is given. For instance, in material and methods "For data standardization, all DICOM files and metadata were validated for compliance with EUCAIM structure", "The DICOM File Integrity Checker developed by our group" → is this material published elsewhere or made open-source through a github/github or similar repository? The amount of work that has been developed here has an incredible value for the community and should be made public whenever possible.
Authors: Thank you for pointing this out. Additional references have been incorporated on page 6 to clarify the availability of the tools and materials mentioned.

- DICOM File Integrity Checker: The DICOM File Integrity Checker developed by the group is not a public tool; however, it is accessible to EUCAIM users for data preprocessing within the project. The manuscript now includes a reference to its entry in the EUCAIM bio.tools catalogue: https://bio.tools/dicom_file_integrity_checker_by_gibi230
- EUCAIM Handbook: A reference to the EUCAIM Handbook has been added. This resource provides a detailed description of the data standardization workflow and the tools available within the project: https://eucaim.gitbook.io/handbook

Reviewer 3.4: Pseudonymisation versus Anonymisation. I have a question regarding this, since both are mentioned during the text, it is not clear to me if the data uploaded is pseudonymised, and there is a log file somewhere (in the hospital submitting the data) that keeps a link (e.g. PatientID) between the data submitted and the real patient identity, or the data is fully anonymised. If the data is totally anonymised and there is no log-file, how do we prevent errors in the future if the hospital submit the same patient (e.g. submission in 2025 for project X, new patient ID projectX_01 - submission in 2026 of same patient for project Y, new patient ID project Y_01). Also, besides removing DICOM tags, can you describe if any action on the images has been done or not? There are some people that advocate to remove facial information from head-and-neck scans, but this can affect the potential use of the data (e.g. in head-and-neck cancer patients). **Authors:** To address this comment, we have included the following explanation and sentences on The Data extraction and exposure pipeline (page 6): "The data preparation workflow includes a two-stage privacy-preserving process:

1. Local pseudonymisation:

Pseudonymisation is performed on a restricted-access Virtual Machine by authorized personnel of the Experimental Radiology and Imaging Biomarkers Platform (PREBI) using an in-house tool developed by the Biomedical Imaging Research Group (GIBI230). The process includes: replacement of PatientID, PatientName and AccessionNumber with specific pseudonym and hashes using Blake2b, renaming of folder structures to remove personally identifiable information (PII), removal of public and private DICOM metadata that may contain PII, exclusion of screenshots (ImageType = SCREEN SAVE) and manual review of

secondary captures (DERIVED or SECONDARY). This step ensures consistent pseudonymised identifiers for repeated submissions of the same patient at the hospital level. The mapping between original and pseudonymised identifiers is maintained only temporarily by the hospital IT service and is not accessible externally.

1. EUCAIM anonymization:

Once pseudonymised data are transferred to the Pseudonymised Medical Imaging Repository, the EUCAIM anonymization pipeline is applied. This pipeline enforces the EUCAIM DICOM Anonymization Profile and generates a unique hash per patient, linked to the project and site, with no retained traceability. Sensitive DICOM headers are removed, and OCR-based pixel-level text detection can be applied to eliminate "burned-in" personal information. File integrity checks and pixel-level duplicate detection can be performed to prevent repeated inclusion of the same images across different projects." We have also reorganized this section adding the following sentence: "This two-layer anonymization approach mitigates re-identification risks, aligning with procedures recommended by the Spanish Data Protection Agency and best practices from the Singaporean authority (Figure 1).

https://openreseurope-files.f1000.com/linked/263811.image_1.gif

Figure 1. Steps for anonymization. Source: Adapted from AEPD. Guide to basic anonymization. Prepared by the National Data Protection Authority of Singapore (PDPC - Personal Data Protection Commission Singapore. Additionally, after transferring it to Pseudonymised Medical Imaging Repository and anonymising the datasets, the DICOM File Integrity Checker developed by our group was used to detect corrupted or missing files. Data were finally ingested into EUCAIM by transferring them to the Reference Node through the QP-Insights API, or by sharing them through a federated node.. For data standardization, all DICOM files and metadata were validated for compliance with EUCAIM's structure and interoperability standards. The data preparation process and the tools involved are defined in the EUCAIM Handbook. Additionally, the metadata of all datasets were registered in the EUCAIM Public Catalogue, which follows the Health DCAT-AP standard and ensures compliance with FAIR principles at the dataset level. An additional layer of dataset discoverability is provided to EUCAIM Data Users through the Federated Query tool, which enables them to perform queries based on specific criteria and retrieve the number of cases (as aggregated numerical results) that meet those criteria"

Reviewer 3.5: Data quality and documentation. You do mention DICOM compliance, anonymisation, etc, but what about data quality and documentation? About data quality, is there any metadata or any quality check to measure the quality of the submitted data? I am thinking more on the annotations of the data, for instance, if the annotations have been done by experts and following consensus international guidelines, etc A concrete example would be the adherence to contouring guidelines in radiotherapy images. Data annotations not being compliant to these guidelines and used to train models might carry future problems. Also, we could think about data quality of the images, if a hospital submit images with very low resolution, artefacts, etc. A way to go would be to enable a community rating for each dataset ... About data documentation. I am putting myself on the user side, and the first question that comes to my mind is: how do I find and decide which data I use for my

project? Having data documentation is very important, like a Data Sheet (similar to what they have done in https://arxiv.org/abs/1803.09010, https://datanutrition.org/, ...). Is there any similar documentation standard in EUCAIM?
Authors: We have reviewed your relevant comment about data quality and we would like to add that efforts are ongoing to define and implement European-level guidelines for data quality, such as those being developed within the QUANTUM project. At the time of dataset preprocessing for this manuscript, these guidelines were not yet applied. In addition, EUCAIM Data Users have access to other tools and workflows to assess data quality. We have included the following comment on the manuscript on page 8: "This rigorous ETL pipeline ensures scalability, data quality through the access to tools and workflows within EUCAIM" In relation with data augmentation, we would like to confirm that EUCAIM provides dataset metadata in the Public Catalogue, which follows the Health DCAT-AP standard and FAIR principles. Researchers can explore datasets and evaluate suitability using the Federated Query tool, which returns aggregated counts for specific criteria, supporting informed dataset selection. We have included the following comment on the manuscript on page 6:" The data preparation process and the tools involved are defined in the EUCAIM Handbook. Additionally, the metadata of all datasets were registered in the EUCAIM Public Catalogue, which follows the Health DCAT-AP standard and ensures compliance with FAIR principles at the dataset level. An additional layer of dataset discoverability is provided to EUCAIM Data Users through the Federated Query tool, which enables them to perform queries based on specific criteria and retrieve the number of cases (as aggregated numerical results) that meet those criteria"

Competing Interests: No competing interests were disclosed.

Reviewer Report 11 November 2025

https://doi.org/10.21956/openreseurope.22734.r63504

© 2025 Strzelecki M. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Michal Strzelecki
Lodz University of Technology, Lodz, Poland

The paper addresses a highly significant issue concerning access to medical imaging data. Such data are essential for the development and implementation of AI algorithms that support diagnostic imaging. Through the EUCAIM project, an effective method for accessing these data has been devised, ensuring compliance with extremely stringent legal and ethical requirements while overcoming technical challenges—primarily related to incompatibilities among hospital PACS systems.
The case study described demonstrates that access to such data is indeed feasible, although at this stage it encompasses only approximately 12,000 images. However, the authors' conclusions are not optimistic. They anticipate difficulties in scaling the developed data-sharing process for

broader clinical implementation, due to diverse data management models within medical institutions and legislative discrepancies governing access to medical data.
As a result, narrowing the gap in the development of such algorithms between Europe and North America or certain Asian countries may become increasingly challenging. Nevertheless, the reviewed work is of considerable importance for the advancement of diagnostic imaging, and its indexing is strongly recommend.

Is the background of the case's history and progression described in sufficient detail?
Yes

Is the work clearly and accurately presented and does it cite the current literature?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: medical imaging, AI, machine learning

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 11 November 2025

https://doi.org/10.21956/openreseurope.22734.r63501

© 2025 Neha N et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Deepak Kumar Shukla
Rutgers Business School, Rutgers University Newark Business School (Ringgold ID: 33869), Newark, Jersey, USA
Neha Neha
Kent State University, Kent, USA

This article documents the implementation of the European Federation for Cancer Images (EUCAIM) initiative, focusing on the integration of imaging and clinical data from a reference university hospital into the EUCAIM infrastructure.

Good points are:
1. High data-processing throughput (12,484 studies, 98.6% efficiency).
2. Clarity and structure suitable for replication within similar EU projects.
3. Integrates lessons from multiple large-scale initiatives.

However Recommendations for Revision:
There is minimal discussion on computational performance metrics or storage requirements. Also;
1. provide pseudonymization scripts, metadata dictionaries, or anonymized examples through Zenodo or EUCAIM's public repositories.
2. include workflow diagrams or tabular summaries of ethical and legal documentation steps.
3. address sustainability models and future interoperability challenges beyond current pilot hospitals.
3. ensure consistent acronym expansion.

References
1. Neha F, Bhati D, Shukla D: Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI*. 2025; **6** (9). Publisher Full Text

Is the background of the case's history and progression described in sufficient detail?
Yes

Is the work clearly and accurately presented and does it cite the current literature?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Artificial IntelligenceComputer VisionImage ProcessingDeep learningMedical Imaging

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have

significant reservations, as outlined above.