# **Repositories for Cancer** Imaging

Ignacio Blanquer Universitat Politècnica de València Institute of Instrumentation for Molecular Imaging



Instituto de Instrumentación para Imagen Molecular





JNIVERSITAT

## Outline

- Motivation and challenges
- The PRIMAGE and CHAIMELEON projects
- The AI4HI Network
- Towards a pan-European Federation: EUCAIM



## **Motivation**

#### ●●● M Al for Medical Imaging Boosts × + ← → C △ @ healthitanalytics.com/news/si-for-medi... ♀ ☆ ☆ ◙ ● & ★ □ @ ANALYTICS IN ACTION NEWS

#### Al for Medical Imaging Boosts Cancer Screenings with Provider Aid

New research shows that a decision-referral approach in which radiologists and artificial intelligence work together for breast cancer screening is better than either's performance alone.



https://healthitanalytics.com/news/ai-for-medical-imagin g-boosts-cancer-screenings-with-provider-aid

- A success story of an AI model trained with 1,193,197 Digital Mammographies
  - In the Valencian Screening program, ~750.000
     women are cited each two years.
  - It will take less than 4 years to create such a dataset.

## **Motivation**

#### ●●● M Al for Medical Imaging Boosts x + ← → C ☆ m healthitanalytics.com/hews/ai-for-medi... Q ☆ ☆ M @ & ★ □ @ ANALYTICS IN ACTION NEWS

#### Al for Medical Imaging Boosts Cancer Screenings with Provider Aid

New research shows that a decision-referral approach in which radiologists and artificial intelligence work together for breast cancer screening is better than either's performance alone.



https://healthitanalytics.com/news/ai-for-medical-imagin g-boosts-cancer-screenings-with-provider-aid

- A success story of an AI model trained with 1,193,197 Digital Mammographies
  - In the Valencian Screening program, ~750.000
     women are cited each two years.
  - It will take less than 4 years to create such a dataset.
- DIPG is one of the deadliest brain tumours in children
  - With 1,5 cases per 100.000 inhabitants per year.
  - We would need 170.000 years.
- Cancer imaging datasets can be a world-scale challenge.

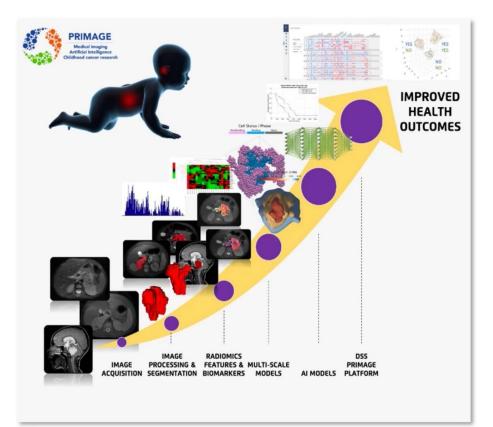
## Motivation and Challenges

- The development of advanced methods for Cancer diagnosis, prognosis and treatment requires large, homogeneous and high-quality imaging data.
- Medical Imaging data comes from Clinical Trials but mainly from Real World Data.
- Collecting a representative annotated dataset is complex due to multiple dimensions:
  - The legal constraints on the secondary use of Medical Data.
  - The risk of patient reidentification.
  - The heterogeneity of multicentre datasources.
  - The reluctance of providers to share data.
- Cancer Imaging Repositories aim at reducing the barriers for those challenges.

## PRIMAGE

### **PRIMAGE** Project

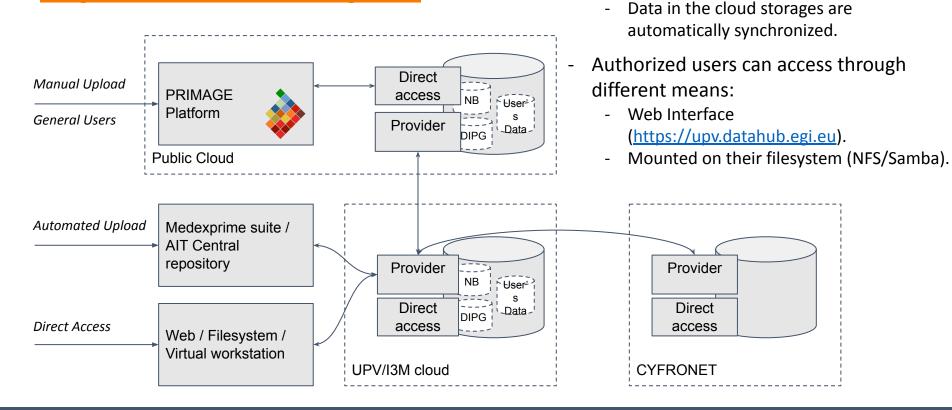
- The main of PRIMAGE is to build an open cloud-based platform as a Decision Support System (DSS) to assist in the clinical management of two paediatric cancers:
  - Neuroblastoma (NB), as the most frequent solid cancer of early childhood, and
  - Diffuse Intrinsic Pontine Glioma (DIPG), being the leading cause of death related to brain tumours in children.
- The PRIMAGE platform implements the latest advancements on imaging biomarkers and tumour growth modelling aimed at assisting clinicians in the diagnosis, prognosis and follow-up of personalized therapy.





#### Data Backend

#### Integration with the Data Management

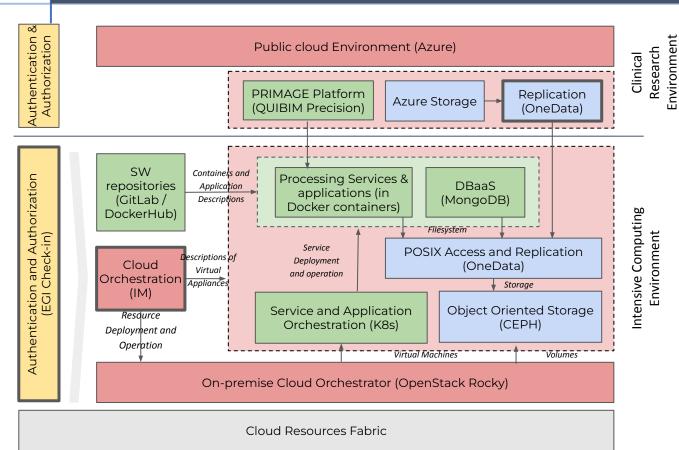


-



Data is uploaded through a web

#### Final Cloud Architecture (gittab)



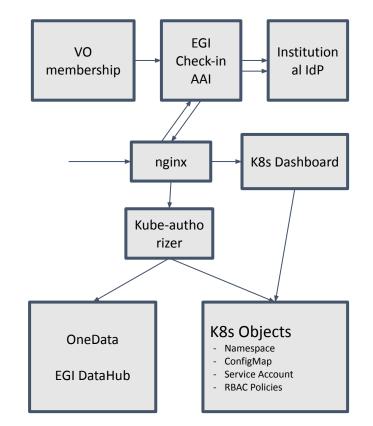
The architecture splits between the Intensive Computing research environment, where AI models training and simulation take place and the Clinical Research Environment where those models are usable for medical researchers.

- The Cloud backend is platform agnostic and can be deployed in different cloud providers.
- The Cloud backend comprises an Object-Oriented Storage and a Service Orchestrator.
- The Service Orchestrator deploys and manages the applications available in trusted repositories.



### Extending the AAI VO model through Kubernetes and EGIDataHub

- A VO has been created to manage access permissions at the level of the Kubernetes (K8s).
- The K8s Dashboard is put behind a proxy that authenticates via EGI-Checkin
  - Only authenticated users that belong to the vo.primage.eu can access the resources.
  - First time a user connects to the Dashboard, a service (Kube-Authorizer) triggers the creation of a set of objects
    - In K8s A namespace, a configmap with the generic configuration, a service account and RBAC policies.
    - Users can only create and browse objects within his/her namespace except for a few specific objects in the default namespace.
    - In EGI DataHub OneDAta, runs a container that creates a "home" directory and a set of permissions.
  - Further accesses go directly to the K8s Dashboard.
- This provides a seamlessly management of multiple sites with the same permission schema
  - Focusing on Repository Federation.





#### Integration Activities

#### **Processing backend**

🗧 😑 🌑 😗 Kubernetes Dashboard	× + ~						
$\leftarrow \rightarrow$ C $\triangle$	u:31443/#/create?namespace=iblanque-at-dsic-dot-upv-dot-es 🔍 a 🖄 🕁 🔊 😔 🗯 🌒 🗄						
🛞 kubernetes	iblanque-at-dsic • Q Search +						
≡ Create							
Cluster							
Cluster Roles	Create from input Create from file Create from form						
Namespaces	Enter YAML or JSON content specifying the resources to create to						
Nodes	the currently selected namespace. Learn more C						
Persistent Volumes	1 apjVersion: v1						
Service Accounts	2 kind: Pod 3 metadata:						
Storage Classes	<pre>3 metadatat 4 name: desktop-tensorflow 5 # Please change the following line with your namespace.</pre>						
Workloads 📧	6 namespace: iblangue-at-dsic-dot-upv-dot-es 7 * spec:						
Cron Jobs	8 restartPolicy: Never 9 containers:						
Daemon Sets	10 - name: desktop-tensorflow 11 image: primage/interactive:ub1804 one2002 tf2						
Deployments	.3.1_gpu_desktop_vnc_sshd						
Jobs	12 imagePullPolicy: Always 13 - securityContext:						
Pods	14 privileged: true 15 · ports:						
Replica Sets	16 - containerPort: 22 17 name: sshd						
Replication Controllers	18 - containerPort: 5900 19 name: vnc						
Replication controllers	20 env:						

→ C △	u:31443/#/pod?namespace=iblang	ue-at-dsic-dot-upv-	dot-es 🔍	ô 🌣 🔝	• *	\varTheta i	$\leftarrow \rightarrow$	C Ó è u
🛞 kubernetes	iblanque-at-dsic	₹ Q	Search		+	<b>\$</b> 0	۲	kubernet
$\equiv$ Workloads > Pods							=	
Cluster Cluster Roles	Pods						Clus	ter uster Roles
Namespaces	Name	Labels	Node Status	Restarts CP	J Usage	(cores)	Na	mespaces
Nodes	console		vnode- 1.localde Running	18 -			No	ides
Persistent Volumes	-	controller-ui	1.locald			- 1	Pe	rsistent Volum
Service Accounts N Storage Classes		d: 1678022 8-45fc-4a14						rvice Accounts orage Classes
	primage-	-b322-2036 8d5e48ae	vnode- Termina 1.locald(Comple	Termina: 0 -				
Workloads 🔋	mfkpt	configmap- mfkpt job-name: c reating-prim			- 1		kloads 📧	
Cron Jobs		age-config map				- 1		on Jobs
Daemon Sets						- 2 of 2		emon Sets
Deployments Jobs						_	Jo	ployments
Pods							Po	
Replica Sets								plica Sets
Replication Controllers								plication Contro

• • • Skubernetes Dashboard	× +		
← → C △	1:31443/#/shell/iblanque-at-dsic-dot-upv-t	dot-es/console/ubuntu?na 🔍 🖞 🖞	🔊 🛛 🗯 🚱
🛞 kubernetes	iblanque-at-dsic 🔻	Q Search	+
	console > Shell		
Cluster			
Cluster Roles	Shell in ubuntu	<ul> <li>in console</li> </ul>	
Namespaces	root@console:/#		
Nodes			
Persistent Volumes			
Service Accounts 📧			
Storage Classes			
Workloads Ň			
Cron Jobs			
Daemon Sets			
Deployments			
Jobs			
Jobs Pods			

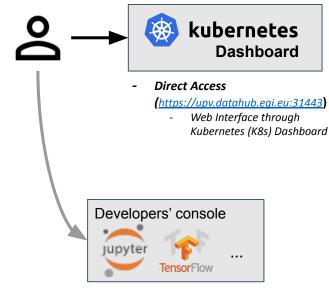
#### Creating applications

#### **Deploying Applications**

#### Accessing Applications

#### Deploying processing jobs through templates

#### Processing backend



- Access to deployed interactive applications:
  - Ports 30050-30070 for Jupyter and similar applications
  - Guacamole web client for desktop applications with SSH and Remote Desktop protocols <u>https://upv.datahub.eai.eu:32443/auacamole/</u>



- Build and Run PRIMAGE Applications (TR2.2 & D2.5 Documents)

https://gitlab.com/primageproject/documentation

- Types of applications:
  - Batch

-

- *High Throughput Computing (HTC)*
- Interactive
- Templates and canonical examples: https://gitlab.com/primageproject/applications
- Docker Containers (DockerHub PRIMAGE Organisation)

https://hub.docker.com/u/primage

- 15 repositories, 42 containers.
- API (python, Java) (GitLab PRIMAGE Repository)
  - <u>https://gitlab.com/primageproject/api-python-k8s</u>
  - <u>https://gitlab.com/primageproject/api-java-k8s</u>





#### **Further Information**

#### **Relevant Links**

- <u>https://gitlab.com/primageproject</u>
- <u>https://www.linkedin.com/company/primag</u>
   <u>eproject/</u>
  - https://twitter.com/primage\_project



#### The PRIMAGE Backend



Using a previously trained model



#### Training a model

#### https://www.primageproject.eu/

	PRIMAGE Medical imaging Artificial intelligence Childhood cancer research				English Search	~	
Home	PRIMAGE Project	Partners	Publications	News and Events	Contact Us		
Workshop Join the Pl Advanced to support oncology PRIMAGE	RIMAGE Workshop on 23/06, simulation and data technoi t the clinical practice in pedi project represented during t nal Conference on Computal	logies atric he	PRedictive In-silico Mu Analytics to support cz personalized diaGnosi prognosis, Empowered biomarkers	ncer and	PRIMAGE is one of the Li more ambitious Europe projects in mellcal ima intelligence and childho	an research ging, artificial	
The projec participati Internatio	the project ct, financed by the European ing in the consortium and ha nally recognized researches cancer are part of the staff o	as an implem in in-silico te	entation duration of 4 ye	ars.	Tweets from @primage_project Primage Project R SIOPE * @SIOPEurope - S We were honoured to me	<b>9</b> Sep 26	

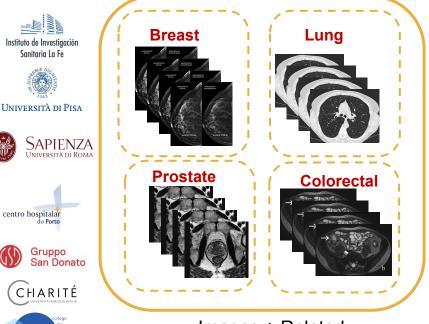




## CHAIMELEON



## **CHAIMELEON Project**



Images + Related clinical data (e-form)

Cloud-based cancer imaging repository as an online resource for the AI community working on the development of cancer management solutions

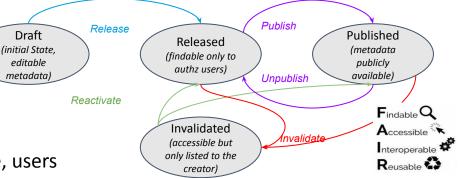
#### Not just a data warehouse...

- Incorporating all necessary functionalities to allow AI experimentation on the cloud (without downloading the data).
- Powered with automation tools.
- Interoperable with other existing initiatives.



#### **Organization of data: A Dataset**

- A CHAIMELEON Dataset is a coherent set of annotated image studies and the associated clinical data that have a persistent identifier.
- A Dataset is a research object that can be citable and fulfils the FAIR principles.
  - Datasets have a metadata that contains aggregated information, following the MIABIS specification, which could be made public (just metadata).
  - Released datasets are discoverable (not necessarily accessible) by the users registered in CHAIMELEON.
  - Published Datasets have their metadata publicly accessible.
- By having the aggregated metadata available, users could raise interest for a specific dataset.
- Access will be granted upon request and after the approval of the Data Access Committee.

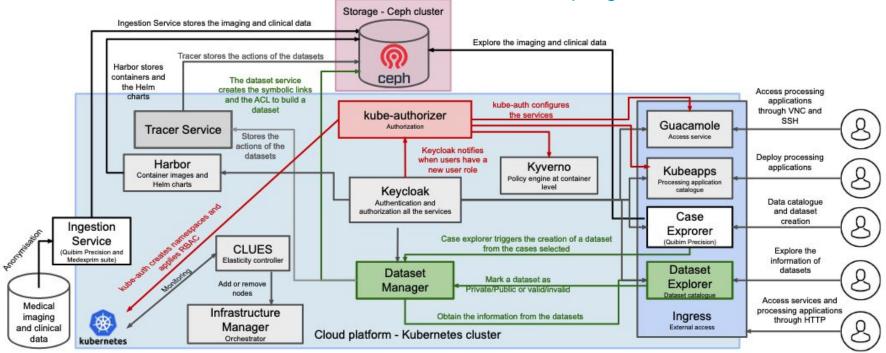


earch and innovation programme under grant agreement No 952172



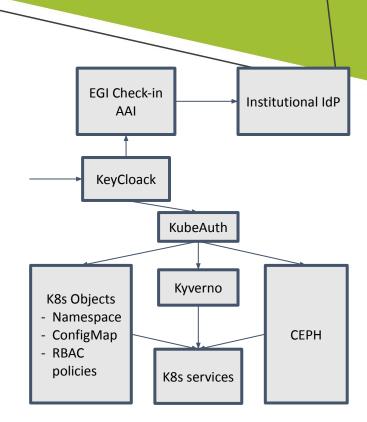
#### **Architecture Technical Set-up**

#### https://doi.org/10.5281/zenodo.6302722 https://github.com/chaimeleon-eu





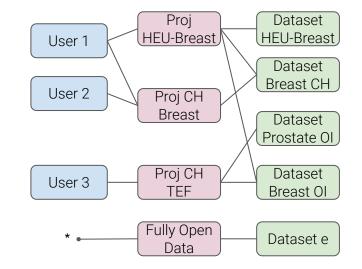
- RBAC in K8s is sometimes insufficient
  - Fine-grain permissions in CEPH, dedicated ingress routes, etc.
- Access to resources may not be only through K8s Dashboard.
- A solution based on KeyCloack and Kyverno has been implemented
  - Kyverno creates the policies for fine-grain permissions.
  - A container creates directly in CEPH the access permissions and the shadow volumes with part of the information
    - Users' home, Read-only permissions for general data and RW for home, etc.
  - Extending the existing functionality of KubeAuth (Namespaces, configmaps and regular RBAC policies).





### **Authentication and Authorisation**

- Authentication is performed through Keycloak<sup>1</sup> and relies on EduGAIN<sup>2</sup> Identity Providers (IdPs)
  - This reduces the burden of managing additional credentials and relies on institutional third party IdP.
- Authorization is based in two concepts
  - Groups represented by projects (tenants)
    - A project has access granted to a set of datasets.
    - Users of a project share the same access permissions.
    - Access to a dataset may be granted to multiple projects.
  - Policies implemented through Kyverno<sup>3</sup>
    - It defines in a declarative way which resources and operations are allowed for a specific project.
  - Access to data is performed through RBAC (Role Based Access Control) policies.



https://kyverno.io/

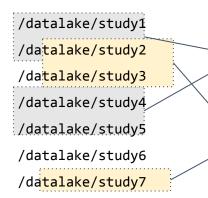
https://www.keycloak.org

//technical.edugain.org/status



### Access model in CHAIMELEON

- Users have access to the data only through controlled virtual environments running on the platform.
- A user can browse and process data in-situ, but cannot download data from the platform
  - Users can upload data and code though.
  - Access is performed through a proxy.
- Datasets are from the data Lake are not replicated, but symbolic links are provided instead
  - Faster, reduces resource wasting and flexible.



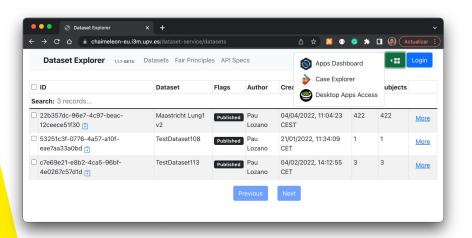
#### Read-Only

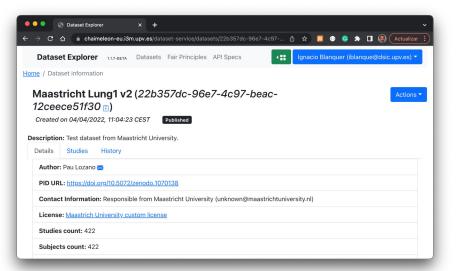
Datasets/Dataset1/link\_study2 Datasets/Dataset1/link\_study3 Datasets/Dataset1/link\_study4 Datasets/Dataset1/link\_study5

Datasets/Dataset1/link\_study2
Datasets/Dataset1/link\_study3
Datasets/Dataset1/link\_study7



#### **Datasets and Metadata**

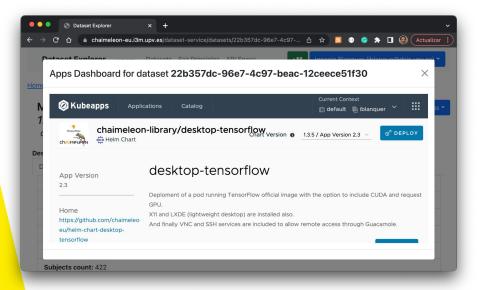




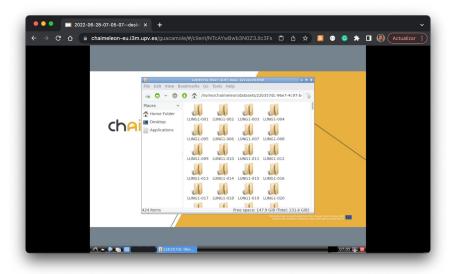
#### https://chaimeleon-eu.i3m.upv.es/dataset-service/datasets



#### Virtual Environments and in-situ access



https://chaimeleon-eu.i3m.upv.es/apps

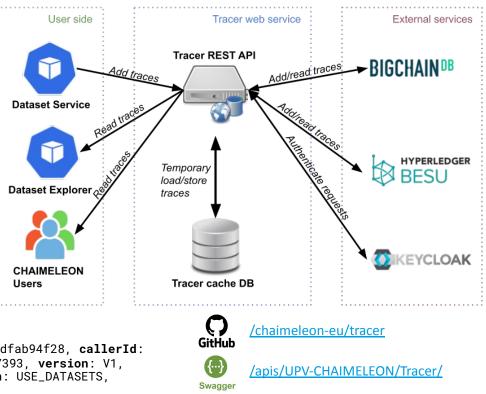


#### https://chaimeleon-eu.i3m.upv.es/guacamole/#/



- It logs user's actions on the cluster (the traces)
  - Create / update (properties of) / use datasets
  - Create / use models
- It stores them in blockchains
  - BlockchainDB
  - Hyperledger Besu
- It does not store user / patient private information
  - Only hashes (for files), ids
- It provides the only entry point to access the blockchain(s)
  - Through Kube network policies
- It ensures traces do not get lost
  - By storing them in cache until the blockchain(s) incorporate(s) them successfully

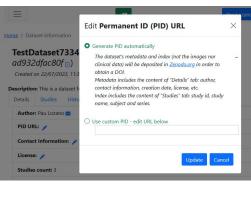
**Use datasets in a pod:** {id: a6dd9a18-c5e0-4fce-bed3-ae7dfab94f28, callerId: e9f5c5cc-cb9c-4c2e-a661-7b6470c7c0da, timestamp: 1663327393, version: V1, userId: f0e9d561-ce6b-44b5-8bdb-2bc9253eea86, userAction: USE\_DATASETS, datasetsIds: [384e1361-09ed-4311-9f19-6bcf9b280fca, d087a6dd-9fce-4b33-82b2-a662bc9ffe47, e9f5c5cc-cb9c-4c2e-a661-7b6470c7c0da]}



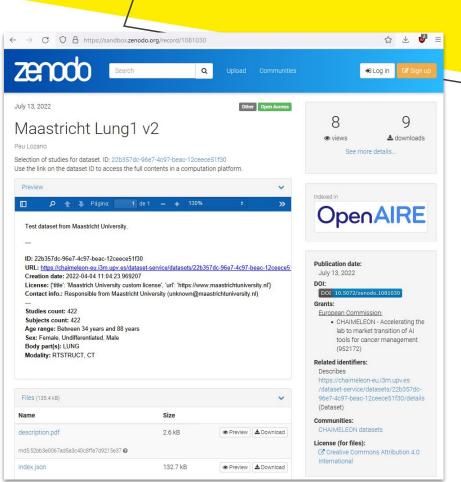


#### Integration with Zenodo

- Dataset service now use the Zenodo API to automatically create a deposition of dataset metadata. This will be done for published datasets and optionally for released datasets.
- The publication at Zenodo gives visibility to the dataset and allow us to obtain a **DOI**, which is a permanent reference (Permanent ID Url) to the dataset that can be included in publications
- Only metadata will be published in Zenodo, never the contents of the dataset (images nor clinical data)
- On invalidated datasets, the Zenodo deposition will be closed (not listed in searches and showing the "closed" label for someone who access through a reference)



through a reference) Currently the Zenodo sandbox endpoint is used, ready to change to production.





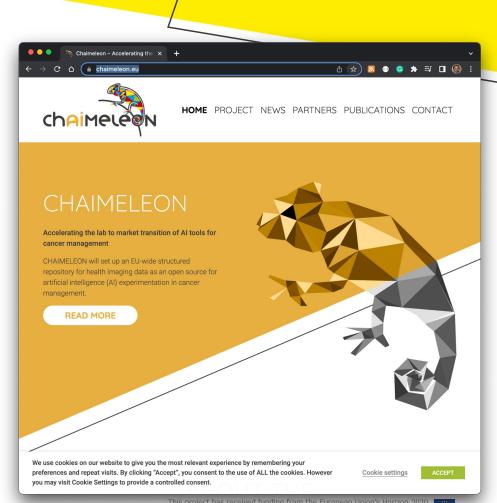


\_

\_

#### **Further Information**

- https://github.com/chaimeleon-eu
- https://chaimeleon.eu/
- <u>https://www.youtube.com/channel/UC</u>
   <u>9jwaCgVs-RpSDIBvXencEQ</u>
  - https://twitter.com/chaimeleon\_eu
  - https://es.linkedin.com/showcase/chai meleon



## AI 4 HI Network

## AI for Health Imaging

- **1. EuCanImage**: A European Cancer Image Platform Linked to Biological and Health Data for Next-Generation Artificial Intelligence and Precision Medicine in Oncology
- 2. **INCISIVE:** 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
- **3. ProCAncer-I:** An AI Platform integrating imaging data and models, supporting precision care through prostate cancer's continuum
- 4. CHAIMELEON: Accelerating the lab to market transition of AI tools for cancer management
- **5. PRIMAGE:** PRedictive In-silico Multiscale Analytics to support cancer personalized diaGnosis and prognosis, Empowered by imaging biomarkers









## AI4HI

- The development of AI models requires a huge amount of high-quality, harmonized data.
- Summing up, the 5 AI4HI imaging repositories cover 5 cancer types (breast, lung, colorectal, prostate cancer, liver and paediatric tumours) over 91,000 patients in total.
- The challenge is to make this data interoperable and accessible
  - o Interoperable goes beyond data formats and metadata.
  - o Accessibility is enormously complex in a context were secondary usage, consent, purpose, data minimization, anonymization and ethic management impose legal obligations and a severe liability.
- The AI4HI Network has been organised into 8 working groups (Ethical and legal issues, Metadata interoperability, Data storage and management, Data annotation, AI development, AI validation, Clinical Working Group and Outreach Working Group), each consisting of 15 experts representing the 5 projects complemented with a wide range of stakeholders, perspectives, approaches and disciplines.











https://www.primageproject.eu

https://chaimeleon.eu https://inc

https://incisive-project.eu

https://www.procancer-i.eu

## **AI4HI Position Papers**

Kondylakis et ol. European Audiology Experimental (2022) 6:29 https://doi.org/10.1186/s41747-022-00281-1 Furopean Radiology Experimenta Onen Access **GUIDELINE/POSITION PAPER** Position of the AI for Health Imaging Check for spdstes (AI4HI) network on metadata models for imaging biobanks Haridimos Kondylakis<sup>1</sup>, Esther Clarrocchi<sup>24</sup>, Leonor Cerda-Alberich<sup>2</sup>, Ioanna Chouvarda<sup>4</sup>, Lauren A. Fromont<sup>5</sup>, Jose Manuel Garcia-Aznar<sup>o</sup>, Varvara Kalolyri<sup>2</sup>, Alexandra Kosvyra<sup>4</sup>, Dawn Walker<sup>7</sup>, Guang Yang<sup>8</sup>, Emanuele Neri<sup>2</sup> and the A4HealthImaging Working Group on metadata models\*\* Abstract A huge amount of imaging data is becoming available worldwide and an incredible range of pr rements can be provided by artificial intelligence algorithms in clinical care for diagnosis and decision support, in this context, it has become essential to properly manage and handle these medical images and to support infrarrowski mad become userial or polycipy intergraves measurements which includes media. And the define which metadata have to be considered, in order for the images to provide their polycimal. All detaids an additional data associated with the images, which provide a complete description of the image exquisition, curation, analysis, and of the relevant chiral available associated with the images. are available to describe one or more subcategories of metadata, but a unique, common, and standard data mod sterableset, to discretion on on here processinglets of modulate double on the standard standard and the admitted of the medical medical transition of the standard all dedicated to the creation of imaging biobanks. Keywords: Artificial Intelligence, Diagnostic Imaging, Metadata, Radiomics, Radiation therap Key points Metadata, as the word suspents, are data about the data · Metadata are essential for the correct use and i.e., additional information about the data themselves. For medical imaging, these include data generated from Metadata are essential for the correct use and interpretation of medical images.
 An appropriate and possibly standardised data an imaging modality, exam prescription codes, descripmodel is necessary to represent these data and their tion data based on an order, and annotations indicating We report the state of the art of metadata models the content and/or anatomy of a particular image [1] Other essential metadata are imaging biomarkers and and the position of Horizon 2020 "AI for Health clinical variables, i.e., complementary non-imaging data Imaging" projects. related to the nationt's medical history that are necessary or a correct diagnosis and decision. In order to efficiently use the medical data, it is crucial to properly combine the actual imaging data with their associated metadata [2]. For this, the appropriate models need to be available to enable homogeneous data access ral Research, University of Riss, Riss, Ital ion is available at the end of the atticle D Springer Open ten body of facilogy 202 Gen Access The article is tenned a loans, which permits us, sharing adaption, distribution and repro popular ones to the original authority and the source, pruvide a lin-so were made. The images or other third party material in this article.

Position of the AI for Health Imaging network on metadata models for imaging biobanks by Haridimos Kondylakis, et. al., European Radiology Experimental Journal. doi.org/10.1186/s41747-022-00281-1 FUTURE-AI: Guiding Principles and Consensus Recommendations for Trustworthy Artificial Intelligence in Medical Imaging by Karim Lekadir, et. al. (2021). arXiv preprint arXiv:2109.09658.

FUTURE-AI: Guiding Principles and Consensus

Recommendations for Trustworthy Artificial Intelligence in

Medical Imaging

Karim Lekadir\*\*, Richard Osuala\*, Catherine Gallin\*, Noussair Lazrak\*, Kaisar Kushibar\*,

Gianna Tsakou<sup>b</sup>, Susanna Ausso<sup>0</sup>, Leonor Cerdá Alberich<sup>4</sup>, Kostas Marias<sup>1</sup>, Manolis Tsiknakis<sup>1</sup> Sara Colantonio<sup>1</sup>, Nickolas Papanikolaou<sup>b</sup>, Zohaib Salahuddin<sup>1</sup>, Henry C Woodruff<sup>1</sup>, Philippe

Lambin<sup>4</sup>, Luis Marti-Bonmati<sup>4</sup>

\*Artificial Intelligence in Medicine Lab (BCN-AIM), Facalty of Mathematics and Compaster Science, University of Barrelows, Spain <sup>1</sup>Marceloil 309, Research and Development Lab. Advans, Genere

- anggori ora, heratita ine zeritati ne zeritati nen zeritati nen zeritati nen orazio, torece "Roudatisi for Research auf Technology, Italiate of Comput Statemen, Greece "Medical Insiging Department, Biomedical Insiging Research Greep, Hospital Universitatio y Bibliotecico La Fe and Health Research Italiati, Builden Comput, David, Builden Comput Statement, Societa Health Research Italiati, Builden Statement, Biomedical David Participation, Sparin Health Research Italiati, Builden Statement, Biomedical David Participation, Sparin Health Research Italiati, Builden Statement, Biomedical David Participation, Sparin Health Research Italiati, Builden Statement, Biomedical David Participation, Sparin Health Research and Participation (Science), Sparin Health Research and Science), Sparin Health Research and Science (Science), Sparin Health Research and

Department of Health, Fundació TIC Salar i Social, Generalitar de Catalanya, Barcelona, Spain

co Abstract

<sup>1</sup>Maastricht University, Department of Precision Medicine, the Netherlands <sup>1</sup>National Research Council, Ionitate of Information Science and Technologies, Italy <sup>1</sup>Champalimand Foundation, Computational Clinical Imaging Group, Lioboa, Portuga

The recent advancements in artificial intelligence (AI) combined with the extensive amount of data generated by today's clinical systems, has led to the development of imaging AI solutions

across the whole value chain of medical imaging, including image reconstruction, medical im-

age segmentation, image-based diagnosis and treatment planning. Notwithstanding the successes

and future potential of AI in medical imaging, many stakeholders are concerned of the potential risks and ethical implications of imaging AI solutions, which are perceived as complex, opaque,

and difficult to comprehend utilise and trust in critical clinical applications. Despite these con-

cerns and risks, there are currently no concrete guidelines and best practices for guiding future AI developments in medical imaging towards increased trust, safety and adoption. To bridge this gap, this paper introduces a careful selection of guiding principles drawn from the accumulated

ences, consensus, and best practices from five large European projects on AI in Health

for specifying, developing, evaluating, and deploying technically, clinically and

Sevenher 30, 2021

impenences, consensus, and best practices from two targe European projects on At in meanin Imaging. These guiding principles are named FUTURE-AI and its building blocks consist of (i)

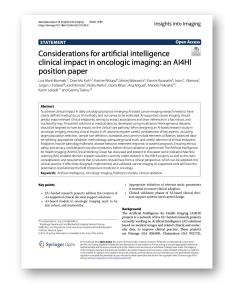
In a step-by-step annoach, these guidelines are further translated into a framework of concrete

Keywords: artificial intelligence, medical imaging, trustworthiness, recommendations, guidelines

Fairness, (ii) Universality, (iii) Traceability, (iv) Usability, (v) Robustness and (vi) Explainab

ethically trustworthy AI solutions into clinical practice

Preprint submitted to arXi



Considerations for Artificial Intelligence Clinical Impact in Oncologic Imaging by Luis Marti-Bonmati, et. al. Insights Imaging (2022). doi.org/10.1186/s13244-022-01220-9



Data Infrastructures for AI in Medical Imaging: A report on the experiences of five EU projects by Haridimos Kondylakis, et al., Journal of Biomedical Informatics

## EUCAIM

## federated EUropean infrastructure for EUCAIM CAncer IMages data

- The EUropean Federation for CAncer IMages (EUCAIM) project originates from an unprecedented body of work and expertise of the "AI for Health Imaging" Network (AI4HI), which consists of 86 affiliated institutions from 20 countries involved in 5 large EU-funded projects on big data and AI in cancer imaging.

-

EUCAIM project is not yet another proposal to build a new infrastructure from scratch, but an integrated architecture carefully designed by the Al4HI Network, and major European Research Infrastructures (Euro-Biolmaging, BBMRI, EATRIS and ELIXIR) on real-world achievements, as detailed in this proposal. The main concept of EUCAIM is of a central hub that federates distributed nodes (repositories, Research Infrastructures, and hospitals) to build up a hybrid distributed and centralized infrastructure on cancer images, including all types of cancer.

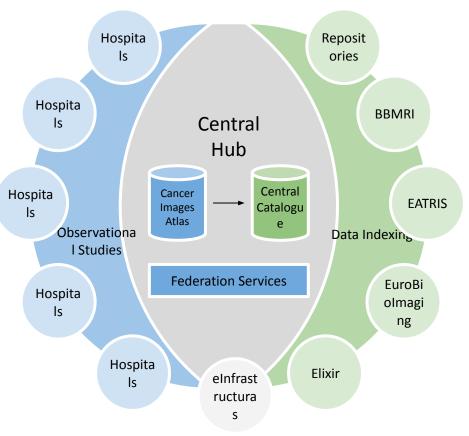
## 79 Partners and 21 Stakeholders

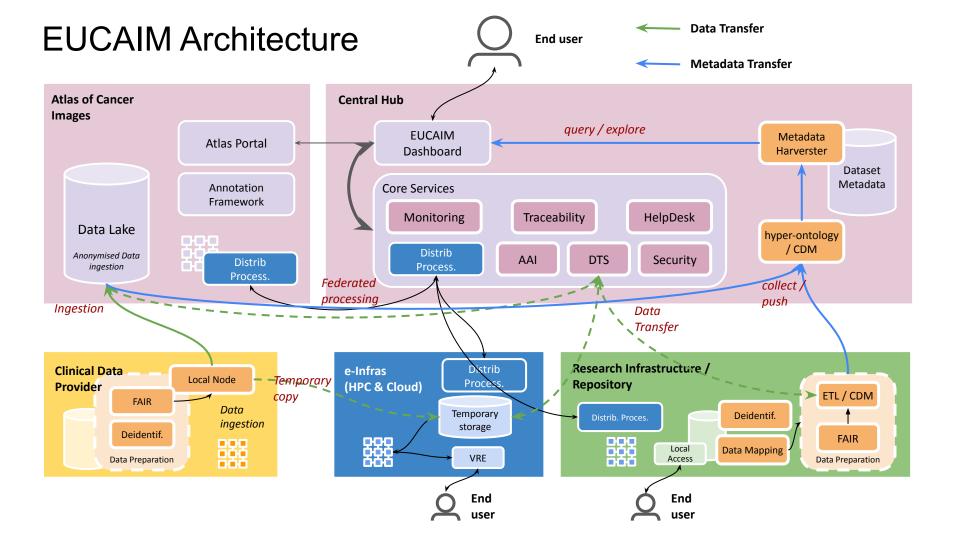




## **Data Scenarios**

- Catalogue of existing Datasets
  - Metadata of available datasets is pushed into a central catalogue (push mode)
  - Datasets metadata is harvested from the providers side.
  - Data is homogenised to follow a Common Data Model.
  - Projects use the existing data as is.
- Observational Studies
  - New data is collected from the research data warehouses at the hospitals.
  - Project-driven approach, data is collected according to the selection criteria.





## Challenges on EUCAIM

- Data Interoperability
  - Data coming from clinical practice and from Observational Studies.
  - Data Preparation and deidentification stages at local providers.
  - Transformation through a Hyperontology and Image Harmonisation.
- Federated Access to Repositories
  - Push / Pull Dataset Metadata Harvester.
  - Data Governance procedures.
  - Redirection and federated access.
- Federated Processing
  - To avoid bulk data transfers and to maximise privacy preservation.
  - In-situ and in trustfull third-party processing.
- Sustainability
  - Recognition models for providers. A complex trade-off between privacy preservation and public recognition.

## Conclusions

- Medical Imaging repositories constitute a great challenge for e-Infrastructures.
- Data privacy preservation together with ethical issues impose non-trivial access restrictions.
- The high computational demand of model training (and even the inference stage for complex Deep Learning models) require the availability of large computing facilities close to data.
- Beyond the FAIR principles, data sovereignty, data interoperability, data harmonisation, traceability and data processing capacity are enormous challenges to tackle the needs of applying AI techniques.

## More information

Ignacio Blanquer

Institute of Instrumentation for Molecular Imaging (I<sub>3</sub>M) Universitat Politècnica de València

iblanque@dsic.upv.es

www.grycap.upv.es

github.com/grycap

chaimeleon.eu/

www.primageproject.eu/

