Interoperability across Inria healthcare software (and beyond)

Researchers perspective

Marco Lorenzi

EPIONE team

...

A variety of solutions/formats

Large heterogeneity and complexity of medical data:

- manual annotations
- medical images
- biological measurements
- clinical assessment
- smart signals

Quality control
Data harmonization
Collaborative analysis

Medical data management and analysis systems are key to guarantee access/governance of sensitive hospital data

We enjoy a **strategical position**

1) long term expertise of several French research groups in analysing medical data
2) strict collaboration of the teams with clinical partners at both national and international level,
3) know-how in software development for medical applications
4) active role in defining and maintaining standards for medical data

Opportunity to set standards and governance for health data analysis

Open and safe management and treatment of healthcare information across the entire life-cycle of medical data

Converge towards shared approaches and standards

-> interoperability between different software solutions

An important coordination effort of research teams and SEDs

We enjoy a **strategical position**

1) long term expertise of several French research groups in analysing medical data
2) strict collaboration of the teams with clinical partners at both national and international level,
3) know-how in software development for medical applications
4) active role in defining and maintaining standards for medical data

Opportunity to set standar

Open and safe managemen **fe-cycle of medical data**

Converge towards shared a

-> interc

An important coordination

MedData4AI

Focus on French research landscape

Enhance hospital participation and contribution

Need for safe, accessible, and open technology

Digital Health Open Source Platforms

Clinica
www.clinica.run
- Quality control
- Data harmonisation

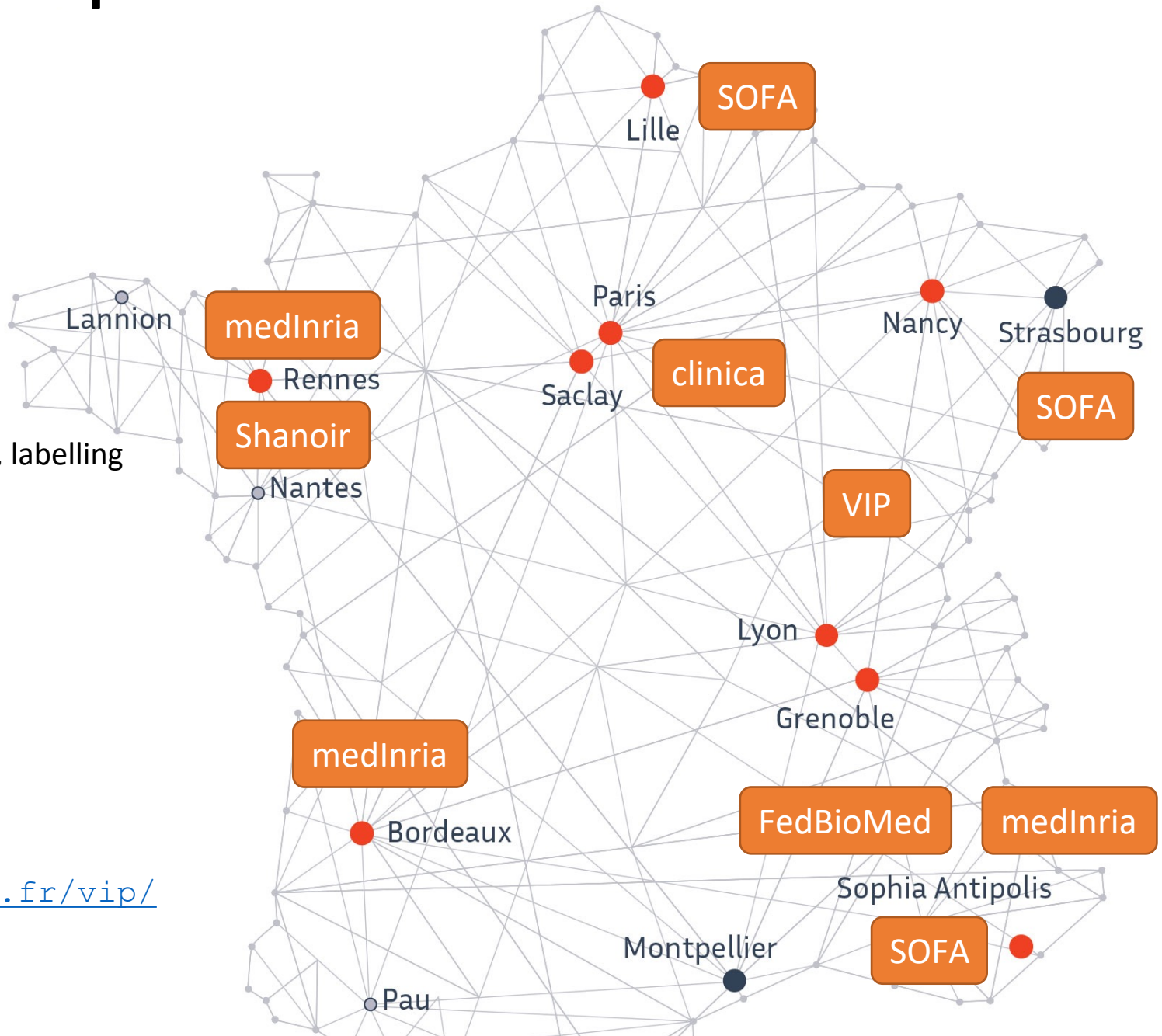
Fed-BioMed
fedbiomed.org
- Collaborative analysis
- Federated learning

medInria
med.inria.fr
- Medical image visualisation, processing, labelling
- AI tools integration

Shanoir
project.inria.fr/shanoir
- Medical image data server
- Microservices

SOFA
www.sofa-framework.org
- Medical simulation
- Real-time interactions

Virtual Imaging Platform
https://www.creatis.insa-lyon.fr/vip/
- Image data analysis



Working group

ARAMIS LAB
BRAIN DATA SCIENCE

Ninon Burgos
Olivier Colliot

SED Rennes
Eric Poiseau

EMPENN

Camille Maumet
Pierre Maurel

SED Lyon
Antoine Fraboulet

liryc
L'institut des maladies du rythme cardiaque

Julien Castelnau

CREATIS

Sorina Pop

Epione
e-patient / e-medicine

Hervé Delingette
Maxime Sermesant

SED Sophia
Marc Vesin

Working rationale

- Defining work-packages around well defined use-cases
- Leveraging existing collaboration networks with hospitals

3 major use cases identified so far (more use cases are welcome ☺)

- EDS AP-HP for brain image analysis
  - 11k patients, MRI scans +clinical data hosted at AP-HP data lake

- EDS AP-HP for prostate cancer analysis
  - 7k patients, multiparametric MRI +clinical data hosted at AP-HP data lake

- Federated learning in Unicancer hospitals (FEDERATED-PET)
  - 1.5k patients, PET-CT images + clinical data hosted by hospitals (decentralized)

Collection of needs through questionnaires for clinicians

- Imaging data modality
- Non-imaging data available
- Annotations
- Analysis tasks
- Data sharing needs

	EDS AP-HP	**Prostate APHP**	**Federated PET**
Modality	T1-weighted & FLAIR MRI	multiparametric MRI (T2w, ADC, DWI, perfusion)	TEP-TDM (3D corps entier)
sample size	Curr. 11k T1 + 13k FLAIR Potentially 130k T1 + 120k FLAIR	5-7k	1k
Raw data	DICOM, Nifti	DICOM	DICOM
Storage	Data lake @ APHP	Data lake @ APHP	Hospital DMZ (local)
Access	VPN	VPN	Local access
Pre-processing	bias field correction, spatial normalisation, tissue segmentation	normalization (B-value) of DWI, rigid registration between sequences, processing of perfusion data to extract dynamic (to be done)	identification des patients éligibles au projet extraction de images TEP-TDM du PACS structuration des dossiers DICOM anonymisation des images conversion Nifti
Software used	Open: Clinica (relies on ANTs, SPM)	Open: for normalization simpleITK	Libre : FedBioMed, scripts python pour conversion des images Nifti et anonymisation Propriétaire : MedXprim, LifeX
Challenges	Quality control: need for manual annotation	quality control aka: Detection of anomalies (artifacts, size, image resolution), standardization of diffusion sequences (selection of the relevant sequence),	Standardisation des images DICOM identiques d'un hôpital à l'autre anonymisation ne détruisant pas les metadata indispensables pour la mesure des SUV (poids, heures d'injection t acquisition…) conversion Nifti respectant les valeurs SUV des pixels avec une interopérabilité d'une machine TEP à l'autre outils de labélisation des images contrôle qualité des images et annotations associées…

	EDS AP-HP	Prostate APHP	Federated PET
Non-imaging data	donneés cliniques/ demographiques/ diagnostiques/ resultats labo	rapports structurés, annotation cliniques	données clinico-biologiques standards de soin courant
sample size	Pour la cohorte de patients avec une T1w : données cliniques pour ~25,000 patients Pour la cohorte de patients avec une FLAIR : données cliniques pour ~80,000 patients	autant que des données d'imagerie (circ 5-7k)	1000
Raw data	Jusque là : tables Hive Maintenant : tables OMOP	.csv (dans le futur peut-etre un schéma indiquant la localisation approximative de chaque lésion)	.csv (dans le futur peut-être FHIR/OSIRIS et donc fichier format json)
Storage	Data lake @ APHP	Data lake @ APHP	données stockées dans un eCRF accessible à distance
Access	VPN	VPN	Chaque centre n'a a accès qu'à ses données. Le data manager a accès aux données de tous les centres
Pre-processing	Appariement avec les données d'imagerie		Si les données sont intégralement saisies dans l'eCRF : pas de pré traitement requis, les ARC ouvrent leur dossier patient informatisé et saisissent directement Si les données proviennent d'une base de données pré-existante : · Soit le centre modifie les données pour coller à l'organisation du CRF cible et l'envoie ensuite - Soit le centre envoie les données « brutes » et le DM réalise le pré-traitement pour pouvoir l'importer
Software used	Aucun en particulier (quelques commandes SQL avant de repasser en Python (pandas)). Il existe une possibilité d'utiliser un outil, Cohorte360		Propriétaire : différents logiciels pour le dossier patient informatisé : Clinicom… / Pour l'eCRF : logiciel EnnovClinical
Challenges	Matcher données cliniques avec une image spécifique (par ex, quel code CIM-10 associer à une image ? Besoin de définir une fenêtre temporelle autour de l'acquisition et voir quels codes ont été attribués dans cet intervalle.)		Si les données sont envoyées brutes, contrôle manuel léger des données avec excel puis transformation et import Si données saisies (ou après import) : programmation des tests grâce au module de test du logiciel EnnovClinical pour envoi de queries aux centres jusqu'à résolution de toutes les incohérences

	EDS AP-HP	**Prostate APHP**	**Federated PET**
Annotations	1- Quality (T1 & FLAIR) 2- Lesion Segmentation (FLAIR)	Annotate the prostate whole gland (WG), the peripherical (PZ) and transition (TZ) zones, the significative lesions	localisation et segmentation des lésions tumorales
Data required	1- Preprocessed {T1w|FLAIR} images, metadata extracted from the dicom header and stored in table (done by EDS people), demographic data 2- Preprocessed FLAIR images	T2w, DWI & ADC sequences	images TEP – TDM
Annotation format	1- Image-level annotation stored in tsv 2- Pixel-wise annotation stored as NIfTI file	json (metadata) + MHA (mask)	Niftii
Time required per annotation	1- ~30s per image 2- ? (future work)	from 5 (only lesion) to 30min (WG+PZ+TZ+Lesion)	20 minutes par images
software	1- Open: Notebook developed specifically for the task 2- ? (future work)	Open: Medinria Proprietary: • MedInria : developed specific plugin that was then made public • Genesis platform from Incepto	Propriétaire : LifeX
Challenges	1- Lack of viewer to define the best annotation setup (e.g. no possibility to scroll through slices to define the best one) 2- Lack of annotation tool	Mechanisms to collect and curate multiple annotations from several raters. Automated quality control . Measure of inter-rater variability	temps d'annotation nécessitant un expert médecin, contrôle qualité des segmentations finales des segmentations des lésions tumorales, stockage des données

	EDS AP-HP	**Prostate APHP**	**Federated PET**
Analysis task	1- Automatic quality control 2- Image-to-image translation 3- Classification for diagnosis prediction 4- Anomaly detection	Lesion segmentation/classification	classification des patients futurs répondeurs versus futurs non-répondeurs
Input Data	Processed {T1w|FLAIR} MRI	multiparametric MRI	données cliniques +images médicales pré-thérapeutiques + données de suivi pour le label: réponse ou non réponse au traitement
Output	1- Quality label, 2- Synthesized image, 3- Diagnosis, 4- Anomaly mask	WG, PZ, TZ, significative lesions, characterization of lesion	Patient label
Algorithms	SVM, DL (CNN, VAE, Unet, cGAN)	DL	DL
Software	Open: Clinica, ClinicaDL, PyTorch	Open:Pytorch	Libre : Keras
Resources required	GPU (32Gb)	CPU-GPU (12-24Gb)	GPU intra-hospitaliers
Challenges	Long computation times Biases caused by imbalances in data quality and heterogeneity	quality assessment, Bias related to diffusion imaging, imaging manufactor, robustness of results.	sensibilité des données médicales, puissance de calcul intra-hospitalière, manque d'ingénieurs hospitaliers ayant les compétences requises

	EDS AP-HP	Prostate APHP	Federated PET
Information to share	QC models: model parameters Image-to-image translation models: model parameters Diagnostic classification models: model parameters Anomaly detection models: model parameters		CNN model parameters, analytics on non-imaging data, Data quality/heterogeneity information Labels
Software			Libre : Fed-BioMed
Privacy Guarantees			On models, statistics, meta-information
Challenges			maintenance suivi et control de l'infrastructure

	EDS AP-HP	Prostate APHP	Federated PET
Other needs	**Visualisation** of data and annotations: visualisation of images and annotation from a remote data lake Handling **heterogeneity** of the data	**Standardization** for certain MR sequences : creation of DWI from high B values, registration between sequences Labelization : **define standard format** for labels, workflow for annotating data before model training but also after applying model for evaluating large cohort, workflow for multiple rater annotation and consensus validation. on test data to pour vérification des résultats et la prise en compte **Quality assessment** and control of data for traning and testing **Visualisation** of data and annotations: visualization of images and annotation from a remote data lake **Modeling data distribution & variability** within a center and between centers, between machine vendors,	Ils sont énormes : les **outils sont hétérogènes d'un hôpital à l'autre,** problème **d'interopérabilité**, **logiciels payants** nécessitant de disposer de budgets hospitaliers spécifiques pour payer les licences annuelles…

WP 1. Towards a General Medical Data Standard for Research

Task 1.1 Extension to non-BIDS compliant modalities.
Liaise with BIDS community and integrate BIDS Extension Proposal (BEP)

Task 1.2 Definition of meta-data format associated to imaging data (e.g. labels, features, post-processed data).
Compatibility with BIDS standard?

Task 1.3 Definition of API for data loading

Task 1.4 Integration with non-imaging data

WP 2. Unified query and retrieval system

Task 2.1. common API for querying imaging + clinical data

Task 2.2. Filtering with respect to quality criteria

Task 2.3. Query with privacy-preserving guarantees

Task 2.4. Retrieving summary information across multiple dataset

Task 2.5. Collecting statistical results / Meta-analysis

WP 3. Towards a unified annotation system for medical imaging data

Task 3.1. Visualization tools, navigation and interactivity

Quality control system (cfr. WP3)

Task 3.2 ML-supported annotations

semi supervised methods, compatibility with ML models format and pipelines

Task 3.3. Secured Remote Annotation of Images

Visualization, annotation, privacy-preserving remote data access/visualization ?

Task 3.4 Collaborative annotation tools + consensus

Centralizing distributed annotations, distributing images to be annotated, labeling through consensus, federated learning

WP 4. Quality control in medical images and clinical data

Task 4.1. Annotation quality control

Cfr WP2.

Task 4.2. Coherence of labels and meta-data

Variability analysis, outliers detection, query syste,m

Task 4.3. Bias detection and removal: imaging and clinical data

Accounting for confounders (supervised), unsupervised analysis, dimensionality reduction

Task 4.4. Visualization and quantification of heterogeneity

Distributed analytics, privacy-preserving methods

WP 5. Collaborative analysis framework

Task 5.1. Performance assessment and results aggregation in multi-centric studies

Task 5.2. Deploying and executing pipelines across multiple participants

Task 5.3. Error handling in multi-centric execution

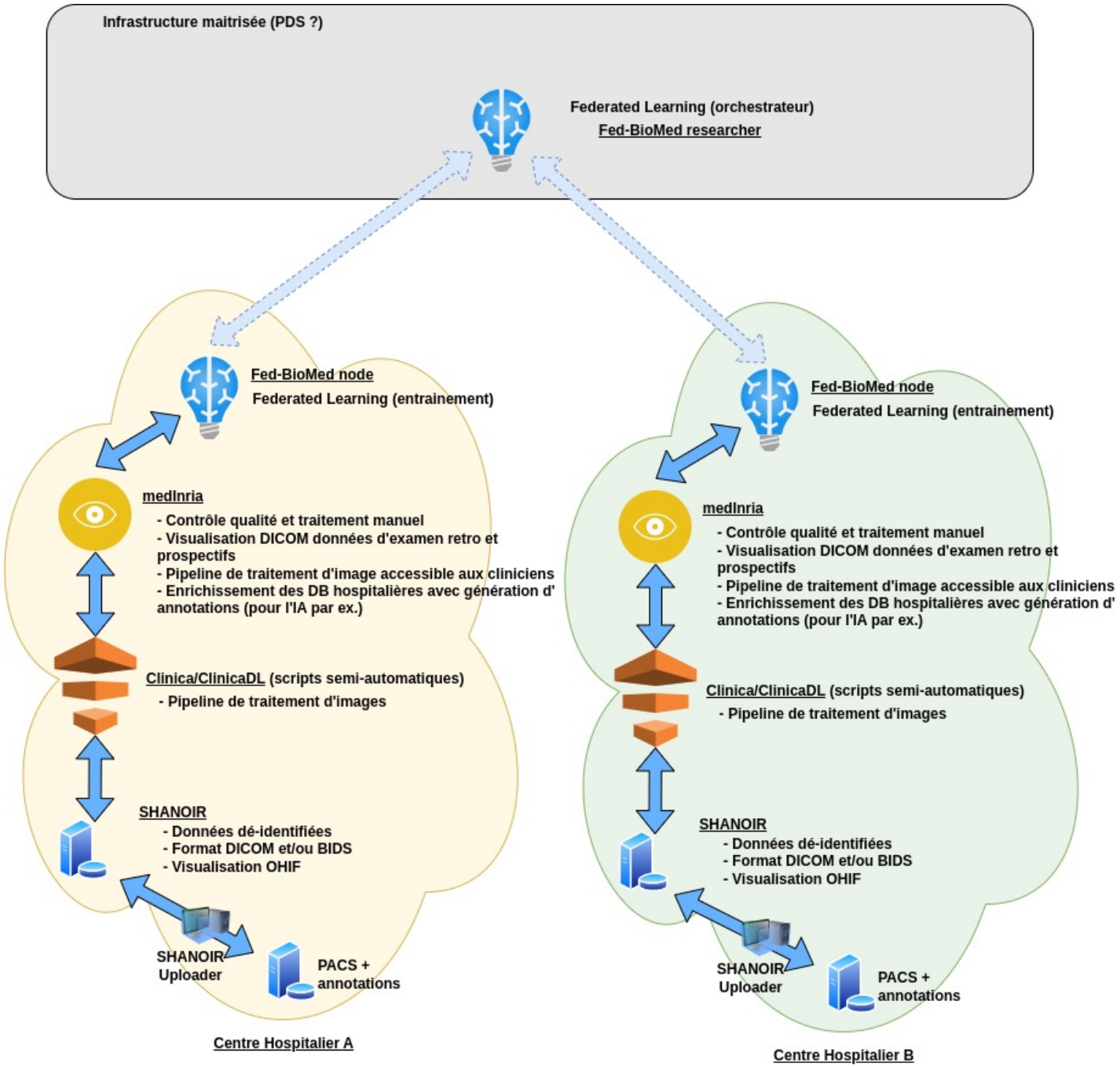
Task 5.4. Privacy guarantees and execution control

WP 6. Enabling Computational Reproducibility by tracking provenance

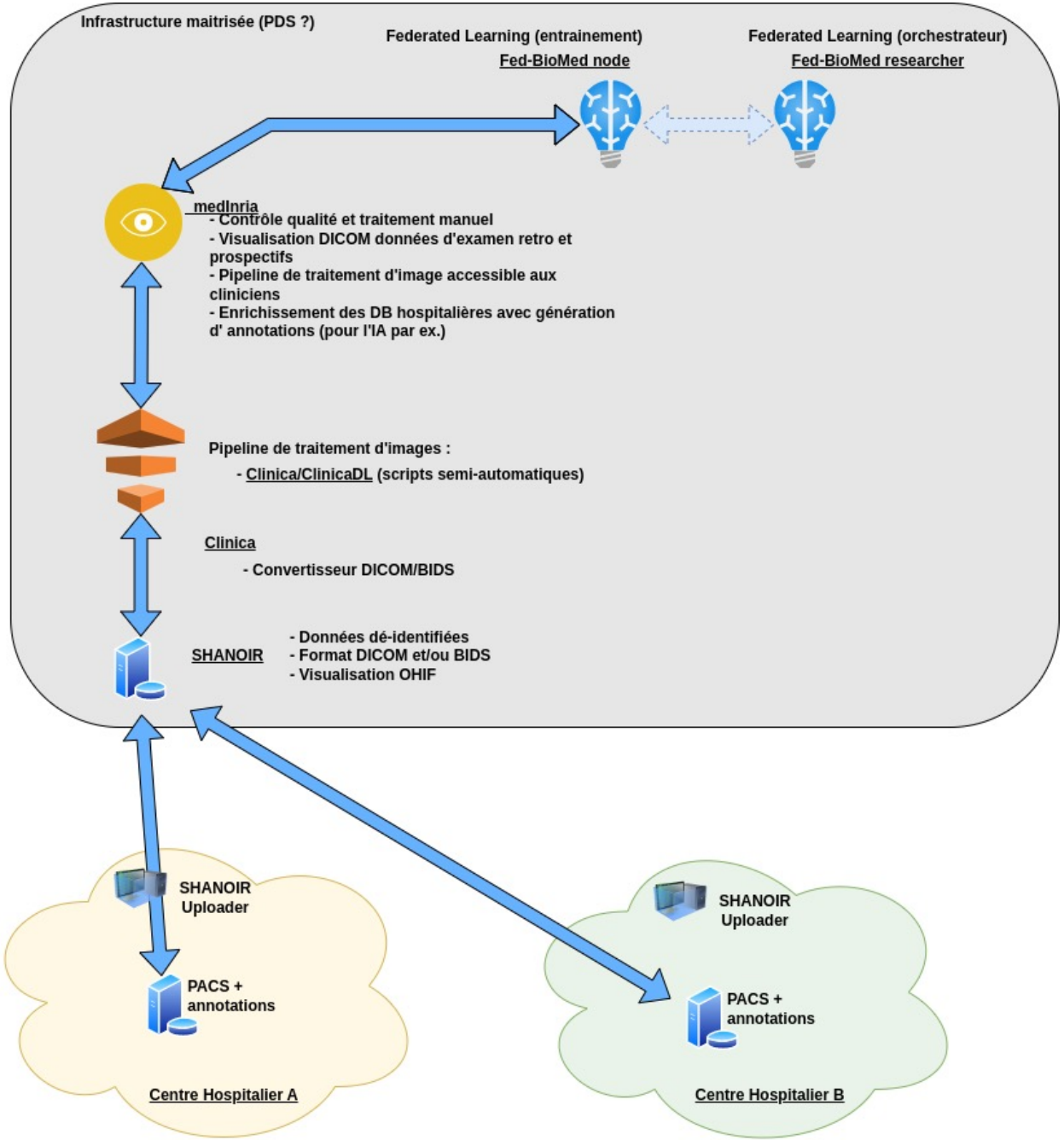
Task 6.1. Implementing provenance standard to store information about how a dataset was processed

Task 6.2. Retrieve and query provenance information

Task 6.3. Use case: comparing results across different analysis pipelines for the different datasets identified?



Thanks to M. Vesin



Thanks to M. Vesin

A structuring program to support current and future research initiatives
PEPR Santé Numerique

Program 1

Axis 1
Methods and models for multimodal multi-scale data integration
Data representation, semantic and structural integration problems, multi-omics, benchmarking
Use-cases from SNDS, imaging-clinical data, -omics

Axis 2
Multi-scale AI for Single-cell-Based Precision Medicine
Integration/standardization of SC datasets, ML for multi-modal data integration, symbolic-AI and ontologies,

Axis 3
Statistical models for longitudinal data analysis
Longitudinal image data analysis

Axis 4
Multi-scale and longitudinal data modeling in pharmacology: towards digital pharmacological twins
Data collection, harmonization, creation federated national database,

Axis 5
Next methodological challenges in clinical trials in the era of digital health
Data integration (IoT, imaging, clinical), integration with EHR, in-silico data

Program 2

Axis 1
Society and Digital Health
Data infrastructure construction to take social determinations into account.
“Platforms of data and quantification must be tackled very urgently“

Axis 3
Traceability for trusted multi-scale data and fight against information leak in daily practices and artificial intelligence systems in healthcare
technical-juridical framework for medical data traceability, certification, patient consent management, resilience to failure of systems, develop platform proof of concept

Axis 4
Secure, safe and fair machine learning for healthcare
Federated learning, bias and heterogeneity, domain adaptation, fairness and explainability

Axis 5
FAIR sharing of reliable protocols to transform datasets into gold standards”
Datasets standardization, annotation and provenance, Architecture for Protocols sharing, Extracting and discovering protocols, Visualization of protocols and datasets, Benchmarking
Neurovascular Use-case

Program 3

Axis 1

Increased multi-scale observability and prediction of cardiovascular diseases

Multi-modal imaging-sensor data,

“Platforms of data and quantification must be tackled very urgently“

Axis 2

“Climatology" of chronic nonischemic cardiomyopathy: Longitudinal multi-modality multicenter analysis and long-time predictive multi-scale modeling

Database creation: imaging, clinical, genetics, biology, electrophysiology and patient data (data lake).

Axis 3

A 5P medicine program to reduce the global impact of stroke

Multi-modal data storage, integration, sharing (IFB, FLI, France Cohortes), FAIR principles.

Program 4

Axis 1

Socially Interactive Agents, behaviors and autonomous digital technologies

Language processing, smartphones, Organize integration of different new technologies, large scale data collection, storage, sharing (biobanks, hospital and cohort's data) and interoperability.

Axis 2

Achieving the interoperability of clinical studies and routine care for the advent of a stratified medicine of aging using clinical, imaging, and omics data

Muti-centric brain data analysis: CATI, federated learning, interoperability, harmonization

Axis 3

High-resolution in vivo and ex vivo imaging data …

Constitution of a Biobank of multimodal and multi-scale brain data, curation of multi-scale/multimodal deep phenotyping biobanks

Networks supported by the France Life Imaging (FLI) infrastructure: REMI and FUN (French Ultra-high field MRI Network)

Axis 4

Quantification of brain health status for better management and early prediction of individual brain health trajectories…

Harmonization of markers for brain image analysis, Integration of BHS tracking workflows with EBRAINS.

Connection with National/International Initiatives?

- Health Data Hub

- CATI

- France Life Imaging

- BIDS: INCF network

- DataLad: distributed data management (https://www.datalad.org/)

- FHIR: standard (https://en.wikipedia.org/wiki/Fast_Healthcare_Interoperability_Resources)

- Federated learning interoperability

- EBRAINS

Conclusions

- Engineering program

- Complementarity with national research initiatives (especially PEPR, Inria Defi)

- Medical imaging focus?

- Pragmatic approach:

    Research/Clinical Needs -> Functionalities -> Milestones -> Resources

- Engineering resources shared between projects (e.g. MedInria <-> Shanoir)

- Centralized coordination (joint with PEPR?)

- Interaction between SED and EPI Inria (and beyond) is fundamental
    Work organization – Roadmap?

Thanks!