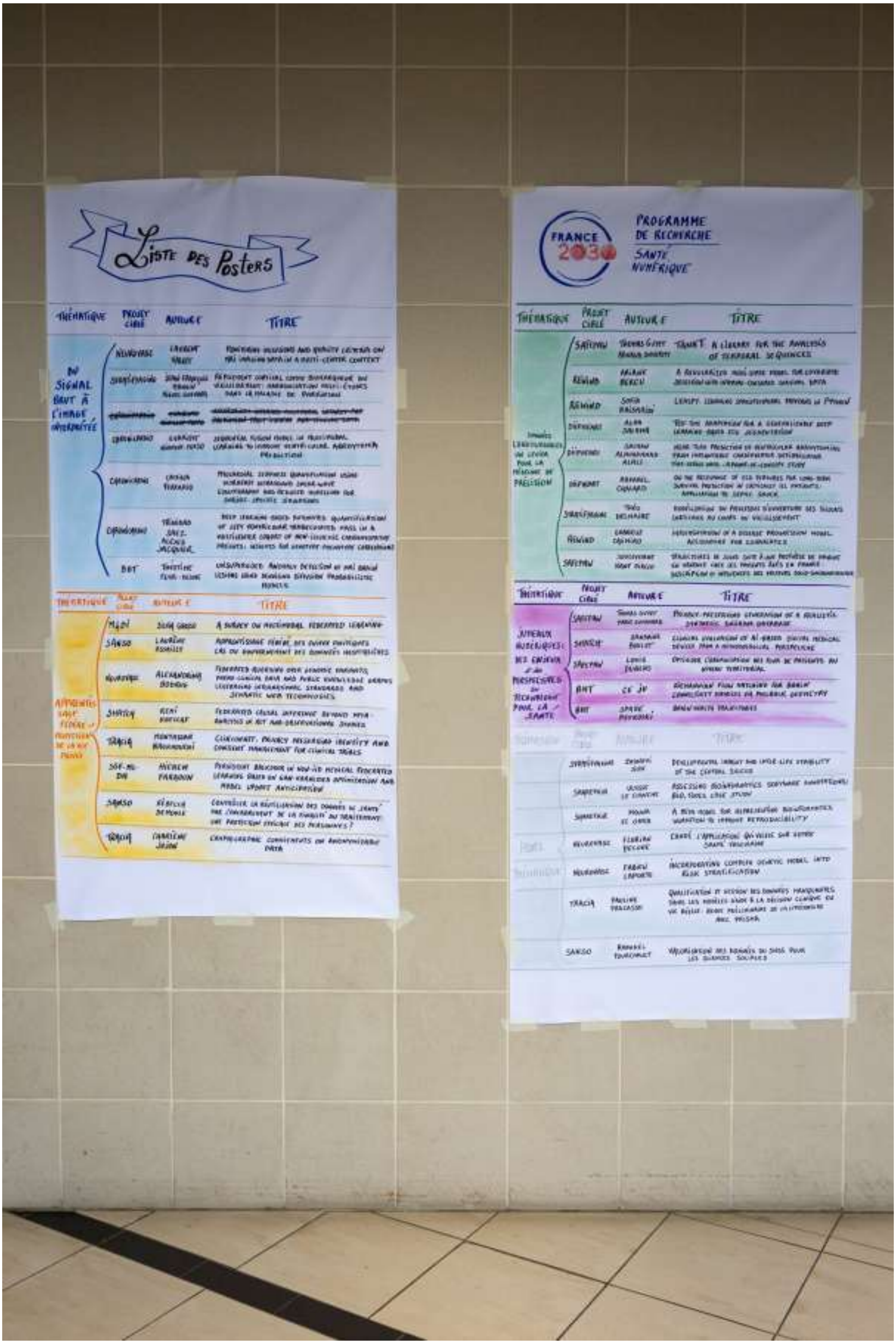


Liste des Posters
THÉMATIQUE
PROJET CIBLÉ
AUTEUR E
TITRE
FRANCE 2030
PROGRAMME DE RECHERCHE
SANTÉ NUMÉRIQUE

Repliement cortical comme biomarqueur du vieillissement : harmonisation multi-études dans la maladie de Parkinson.

M. Guevara[1,2], A. Lanore[1], C. Fischer[2,3], M. Chupin[1,2] et J-F. Mangin[2,3]

[1] Institut du Cerveau et de la Moelle épinière, Paris, France;

[2] CATI, US52-UAR2031, CEA, ICM, Sorbonne Université, CNRS, INSERM, APHP, Ile de France, France

[3] Université Paris-Saclay, CEA, CNRS, Neurospin, Baobab, UMR9027, Gif-sur-Yvette, France.

Background

- **Hausse** des maladies neurodégénératives comme la maladie de **Parkinson** [1].
- La **détection précoce est difficile**, notamment lorsque les symptômes sont discrets [2].
- De **nouveaux biomarqueurs** sont nécessaires au-delà des mesures structurelles classiques [3].
- **Ouverture des sillons** (Fig. 1) testée comme biomarqueur de la maladie de Parkinson sur des **données harmonisées**.

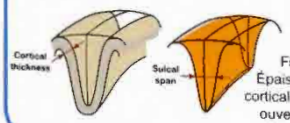


Fig. 1: Épaisseur corticale v/s ouverture sulcale

Méthode

- Données PPMI multi-centriques [4] (Fig 2):
  - 43 batch (centre + protocole)
  - Contrôle qualité des IRM T1.
  - 1609 sujets (V1) : 104 HC / 878 PRodromaux / 591 PD (65,2 ± 8,3 ans; 916 H / 693 F).

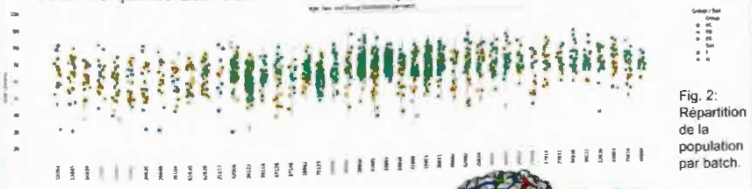


Fig. 2: Répartition de la population par batch.

- Ouverture de **55 groupes de sillons** (Morphologist) [5] (Fig. 3).
- Correction des effets de batch par **ComBat** [6].
- Analyses statistiques : MLM et OLS, avant/après harmonisation, avec ANOVA et post-hoc Tukey [7].

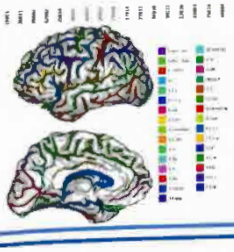


Fig. 3: Groupes de sillons.

Résultats

- Harmonisation: ComBat ≈ OLS / MLM → effets de batch contrôlés (Fig. 4).
  Covariates: Age, sex, groupe.
- Ouverture sulcale :
  - ANOVA: Sillons significativement différents dans l'hémisphère gauche, FDR-corrigé (Fig. 5A):
    - F.C.L.a (scissure latérale antérieure)
    - F.C.L.r (rameaux antérieurs)
    - S.Pe.C.inf (sillon précentral inférieur)
    - Insula
    - S.F.sup (sillon frontal supérieur)
  - Post-hoc Tukey:
    - PD↑/ PR↑ > HC (Fig. 5B)

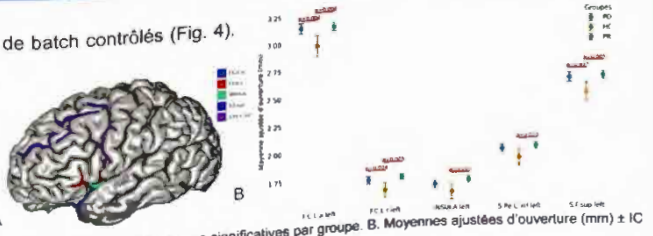


Fig. 5 : A. Sillons avec différences significatives par groupe. B. Moyennes ajustées d'ouverture (mm) ± IC 95%; p-values significatives (Tukey p ≤ 0,05) en rouge.

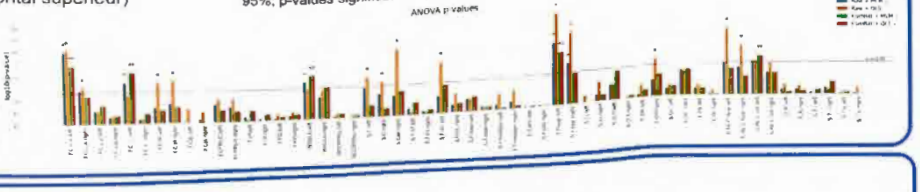


Fig. 4 : Effet du modèle et de ComBat sur l'ANOVA
*Significatif après FDR
✓Effet batch corrigé.

Conclusions

- L'harmonisation est indispensable: **ComBat corrige** efficacement les **effets de batch**.
- L'**ouverture des sillons** apparaît comme un **biomarqueur prometteur** de la maladie de Parkinson.
  - S.F.sup et S.Pe.C.inf augmentent chez PR et PD, suggérant des altérations précoces des régions motrices (ex. l'aire de Broca).
  - PR ≈ PD : changements précoces et stables, possibles indicateurs du sous-groupe à risque chez les PR.
- Perspectives : corréler ces marqueurs à d'autres données cliniques pour évaluer leur valeur prédictive.

Références

[1] Balestrino, R., & Schapira, A. H. (2020). Parkinson disease. European journal of neurology, 27(1), 27-42.
[2] Raby, V. et al (2025). Stability and accuracy of a diagnosis of Parkinson disease over 10 years. Neurology, 104(9), e213499.
[3] Wang, E., et al (2021). Patterns of Sulcal depth and cortical thickness in Parkinson's disease. Brain Imaging and Behavior, 15(5), 2340-2349.
[4] www.ppmi-info.org/data. Données du 27/11/2024.
[5] https://brainvisa.info
[6] Pomponio, R. et al. (2020). Harmonization of ...
[7] ... (lme4 (Bates et al., 2015), emmeans (Lenth, 2021)) and Python (statsmodels (Seabold & Perktold, 2010)).

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0... StablyAging

Laurent VALLET

Monitoring inclusions and quality criteria on MRI imaging data in a multi-center context

Laurent VALLET[1], Solène JOUAN[4], Julien LOUIS[2], Alban GAIGNARD[1,3], Michael KAIN[2], Romain BOURCIER[1,4] and Richard REDON[1]

1. Nantes Université, CHU Nantes, CNRS, INSERM, l'Institut du thorax, F-44000 Nantes, France
2. INRIA, Empenn Team, ERL U1228, F-35000 Rennes, France
3. IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France
4. Neuroradiological Department, University Hospital Center of Nantes, France

FROM UNIVERSITY HOSPITALS TO A LARGE SCALE IMAGING DATASET FOR CLINICAL RESEARCH

Intracranial aneurysms are most often discovered incidentally. In the event of rupture, they cause subarachnoid hemorrhage with a mortality rate of 30–50%[1]. A better understanding of their pathophysiology is crucial to **detect aneurysms, predict their evolution and to optimize therapeutic strategies**. The objective of the PEPR *Santé Numérique* Neurovasc is to set an optimal digital ecosystem to develop predictive tools and models for clinical research on intracranial aneurysms and stroke outcomes. Gathering a group of **29 University hospitals as contributors for the imaging dataset**, we face the challenge of facilitating, filtering and monitoring this data collection to ensure quality and reliability, key factors for the effective training of predictive models.

COLLECTING THE DATA

Imaging examinations are collected during care routine from different hospitals. The diversity of practices and acquisition devices among these contributing centers produces heterogeneous imaging data. This data source, stored in hospitals storage systems called PACS, contains clinical and personal informations at this point.

Collecting those data requires querying the PACS and performing de-identification.

Moreover, a **quality control at import** is required to ensure that only imaging data meeting the usage requirements are uploaded to the research database. It aims at saving storage resources as well as avoiding costly data curation efforts.

THE DICOM STANDARD

The Digital Imaging and Communications in Medicine (DICOM) standard defines both a file format and a communication protocol. It is the worldwide reference for the storage, exchange, and interoperability of medical imaging data.

A DICOM study is organized as follows:

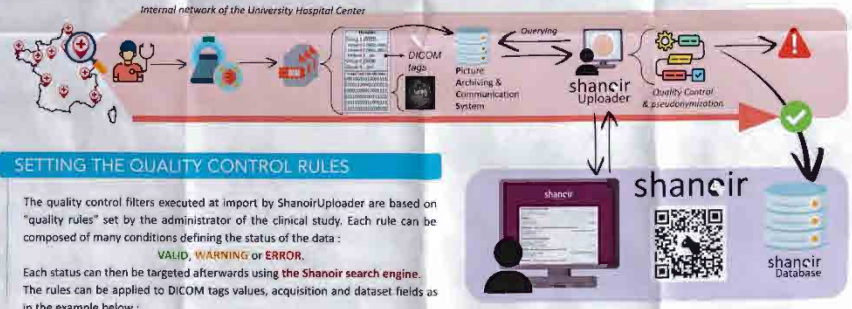
- it contains one folder for each acquisition, with each folder identified by a unique identifier.
- each acquisition folder stores the image data together with **associated metadata**, indexed by standardized identifying codes known as **DICOM tags.**

SHANOIR SOFTWARE

Shanoir[2] is an **open-source database software** developed by INRIA with the resources and support from France Life Imaging.

It is composed of:

- **Shanoir** : a web platform to administer medical imaging data. It **supports DICOM, BIDS, Bruker & EEG formats** and allows clinical study promoters to manage users rights, set participating centers informations, acquisition equipments and other features.
- **ShanoirUploader** : a cross platform client made to **upload imaging data from a PACS** to the Shanoir server after pseudonymization[3]. Shanoir Uploader can run a quality filter at import as described below.



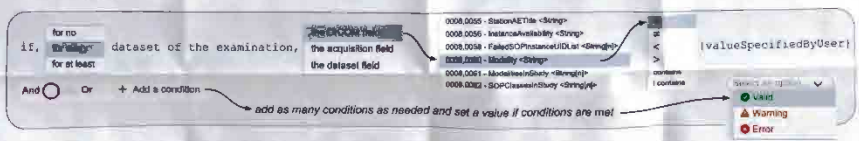
SETTING THE QUALITY CONTROL RULES

The quality control filters executed at import by ShanoirUploader are based on "quality rules" set by the administrator of the clinical study. Each rule can be composed of many conditions defining the status of the data :

VALID, WARNING or ERROR.

Each status can then be targeted afterwards using **the Shanoir search engine.**

The rules can be applied to DICOM tags values, acquisition and dataset fields as in the example below :



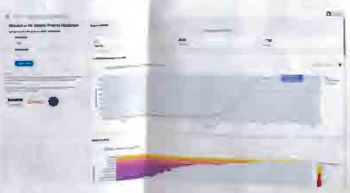
MONITORING THE IMAGING DATASET

In addition to the quality control, we developed a **secured API exposing real-time statistics** from Shanoir database for the benefit of study administrators and contributors.

→ Follow progression of the studies, detect issues or delays

→ Ensure consistency of the collected data

→ Statistics are enriched with JSON-LD annotations to **enhance discoverability**



Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0008

Zhongyi SUN

Developmental Imprint and Later-Life Stability of the Central Sulcus

Z. Y. Sun[1,2], T. Rivière[1], C. Fis[illegible], A. Makin[3,4,5] and J.-F. Mangin[1,2]

[1] [illegible], CEA, Paris-Saclay University, CNRS, Gif-sur-Yvette, France

[2] [illegible]maging.com, UMS2, UAR3raging, ICM, CEA, INSERM, CNRS, APHP, Sorbonne Université, Paris, France

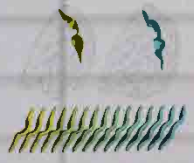
[3] MRC [illegible]ain Sciences Unit, University of Cambridge, Cambridge, UK

[4] Institutes of Cogn[illegible] University College London, London WC1N 3AR, United Kingdom

[5] Wellcome [illegible]for Human Neuroimaging, UCL Institute of Neurology, London, UK

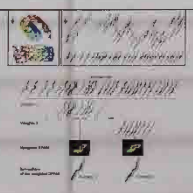
Introduction

The central sulcus (CS) is a reliable anatomical landmark whose morphology and functional organization are sensitive to both developmental influences and age-related change. A key feature along the CS is the *hand knob*, a robust marker of the hand area in motor cortex (Yousry et al., 1997). The CS begins to form around 10–12 weeks of gestation, establishing early scaffolding for [illegible] functional specialization. In aging, widening of the sulcal opening, cortical thinning, and reduced functional specificity in motor and somatosensory cortices are consistently observed. The study of congenital one handers and amputees provides a unique opportunity to examine two key questions: whether sulcal form is sensitive to early developmental deprivation, and whether it remains stable following later-life input loss. Hence, addressing the stability of CS morphology is particularly important, as atypical sulcal patterns in genetic neurodegenerative diseases may reflect subtle developmental alterations long before clinical onset, offering a potential biomarker for early detection and patient stratification.



Methods

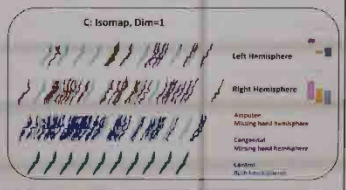
The dataset was composed of three groups: controls (CTR; n=24), congenital one-handers (CONG; n=25), and amputees (AMP; n=16). First, the two central sulci were extracted automatically using the "Morphologist 2012" toolbox of BrainVisa [http://brainvisa.info] and brought to Talairach space with an affine transform. Right sulci were flipped to become comparable with left sulci. Then, the shape similarity between two sulci was coded by the average quadratic distance between their 3D representations after pairwise rigid alignment. Finally, from the resulting similarity matrix, we examined CS morphology using nonlinear dimensionality reduction algorithms: Isomap (Tenenbaum et al 2000) and UMAP (McInnes et al, 2018). Hence, localization of a specific sulcus in the manifold provides a summary of its shape (Sun et al, 2012)

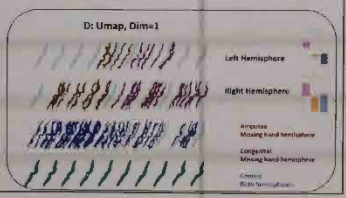


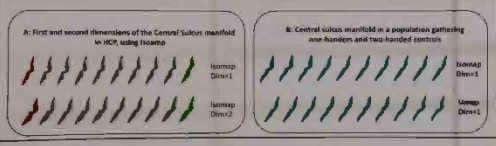
Results

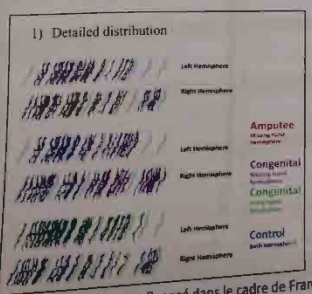
In the hemisphere contralateral to the missing hand, congenital one-handers showed a flatter CS relative to both amputees and controls. This alteration was not observed in amputees. In addition, a distinct CS configuration—a ventrally displaced, curvilinear, and deeply inflected omega-shaped hand knob—was selectively preserved in the right hemisphere of congenital one-handers. There was no shape difference found in the hemisphere ipsilateral to the missing hand. Exploratory analyses further revealed associations between CS shape in congenital one-handers and functional activation of the missing hand area for lips, feet and arm contralateral to the missing hand.

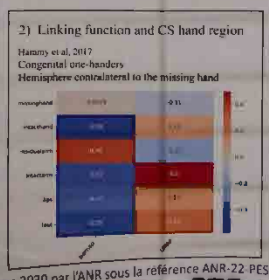
These findings suggest that flatter folding in congenital one-handers reflects the early absence of sensory–motor input, while the preserved morphology in amputees highlights the stability of gross sulcal structure of later-life deprivation.

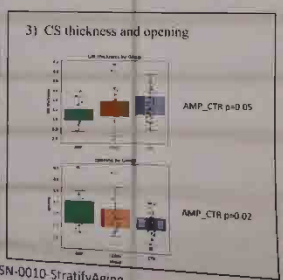












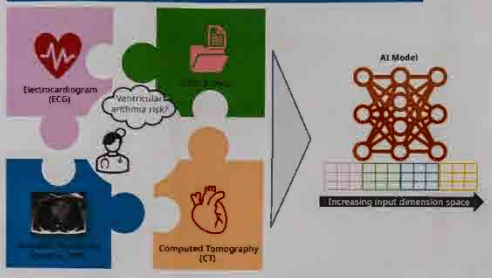
Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0010-StratifyAging

Sequential Fusion
Constraint-Based Model in Multimodal Learning to Improve Ventricular Arrhythmia Prediction

Evariste Njomgue Fotso[1]
Buntheng Ly[2]
Hubert Cochet[2]
Maxime Sermesant[1,2]

1 Inria, Université Côte d'Azur, Epione team, Sophia Antipolis, France | 2 IHU Liryc, Université de Bordeaux, Bordeaux, France

Multimodal classification: problem statement



Motivation

- Capture **complementary** information coming from different modalities in and **increasing dimension** setting.

Definitions

- **Fusion** is the method for combining data from multiple modalities.
- **Constraint-Based** models involves incorporating explicit rules, logical constraints, or domain knowledge into the learning process.

Multimodal classification: State-of-the-art

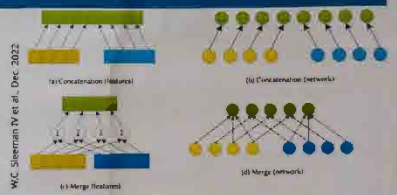


Fig 5: A depiction of the main data fusion techniques (concatenation, merge) with two modalities (yellow, blue) and the final result (green). Traditional machine learning features are depicted as squares, neural network nodes as circles, and feature merging operators as Σ.

Early, Intermediate or Late Fusion

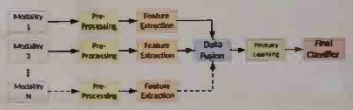
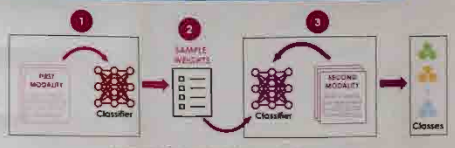


Fig 1: An example architecture for multimodal classification with early fusion.

- At the moment in traditionnal machine learning, early features concatenation is mainly used.
- In deep learning, late network and score concatenation are mainly used

Sequential Fusion: Novel fusion technique for multimodal classification



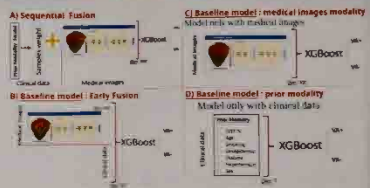
Sequential Fusion Model: general setting

- A first classifier is trained only with the first modality (the prior modality).
- The classification result on the prior modality is used to compute the prior modality constraint (sample weights):
- This constraint is added into the learning process with the second modality to get the final output.

Main idea behind the sample weights (wl) computing

- More importance is given to misclassified samples by the prior modality so that the final classifier is more focused on these samples.
- Yl is the true label of the l-th sample and pl is the predicted probability of the l-th sample belonging to the positive class
- This formulation can be generalised in multi-class classification.

Sequential Fusion: Application for Ventricular Arrhythmia prediction

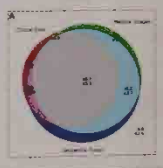


Different model configurations compared

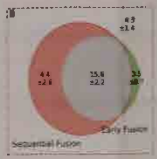
Table 1: Clinical Data Modality | N = 600

Feature Name	Statistics: [min; max] mean or proportion
LVEF [%]	[10; 74] 35.07 ± 13.86
Age (years)	[25; 99] 72.89 ± 12.09
Smoking	270 (45%)
Dyslipidaemia	396 (66%)
Diabetes	159 (27%)
Hypertension	394 (66%)
Sex	100 Females - 500 Males
VA	Positive (VA+) - 185 (28%)

Description: LVEF [%] - Left Ventricular Ejection Fraction, Age - Patient age at the scan date.



A. Venn diagram comparing well-classified samples by each modalities and Sequential Fusion model. Values are means and standard deviations on tests set. The clinical data classifier is red, the medical images classifier is green, and the Sequential Fusion classifier is blue.



B. Venn diagram comparing well-classified positive samples (VA+) by Early Fusion and Sequential Fusion models.

Table 2: Model performance results. Best performance metrics in **bold**.

Model	Accuracy	F1 Score	Sensitivity	Specificity
(A) Sequential Fusion	**0.833**: ±0.04	**0.731**: ±0.06	**0.807**: ±0.04	**0.844**: ±0.06
(B) Early Fusion	0.790: ±0.03	0.631: ±0.06	0.650: ±0.08	**0.844**: ±0.04
(C) Medical Images Modality	0.773: ±0.04	0.607: ±0.05	0.629: ±.06	0.825: ±0.05
(D) Prior modality	0.682: ±0.04	0.521: ±0.06	0.628: ±0.10	0.703: ±0.05

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0015

Myocardial Stiffness Quantification Using Ultrasound Shear Wave Elastography and Reduced Modeling for Subject-Specific Simulations

Camilla Ferrario[1,3], Jairo Rodríguez Padilla[1,3], Maelys Venet[2,3], Olivier Villemain[2,3], Maxime Sermesant[1,3]

[1] Inria - Université Côte d'Azur, Epione Team, Sophia Antipolis, France [2] Bordeaux University Hospital (CHU) - Department of Pediatric and Adult Congenital Cardiology, Pessac, France
[3] Electrophysiology and Heart Modeling Institute Institut Hospital-Universitaire (IHU) -Liryc - Fondation Bordeaux Université, Bordeaux, France

Full paper!

Motivation and clinical context

Myocardial stiffness (**MS**) is the **resistance of the heart muscle to deformation**, combining passive elasticity and active contraction. It regulates filling and ejection and is a hallmark of many cardiac pathologies [2][3].

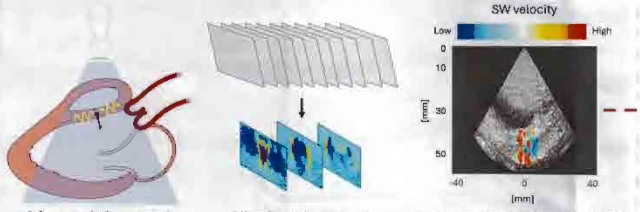
From SWE to personalized heart mechanics

We fit SWE-derived MS curves by optimizing active-stiffness parameters in a personalized 0D cardiac model [1].

Materials and Methods

CLINICAL DATA ACQUISITION

Push (ARF) — Track (UUI) — Quantify (SWE)



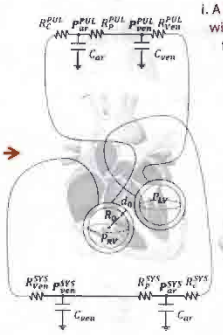
A focused ultrasound pulse generates waves in the septum

Ultrafast ultrasound tracks shear waves (SW) propagation

SW velocity c_s yields MS [2][3]:

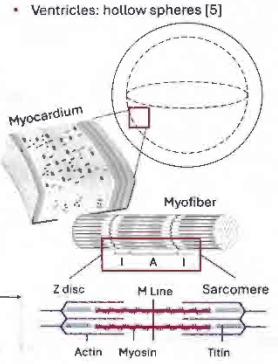
$$\mu \approx \rho c_s^{\,2}$$

CLOSED-LOOP 0D CARDIOVASCULAR MODEL



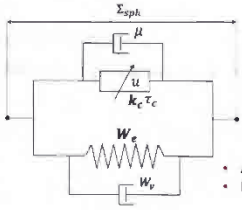
i. A *lumped-parameter* circulation is coupled with a *BCS-based* EM heart [4] to simulate flow and pressure dynamics via ODEs.

- Circulation: *Windkessel* models
- Ventricles: hollow spheres [5]

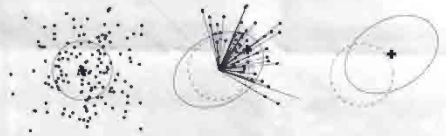


ii. Myocardium as a *Hill–Maxwell* rheological system [5][6]:

- Active cross-bridge (XB) contraction
- Passive hyperelastic/viscous response



OPTIMIZATION USING CMA-ES ALGORITHM



Sampling — Estimation — New distribution

Why?

Stochastic, gradient-free optimizer that adapts its sampling toward promising regions [7]

- CMA-ES estimates maximum stiffness, contraction and relaxation rates (k_0, k_{ATP}, k_{SR}) by minimizing the mismatch between simulated and clinical MS curves.
- To address the *ill-posed parameter identification problem*, we plan to apply the **IUP algorithm** [8], which iteratively updates population priors and guide parameters toward a consistent low-dimensional subspace.

• Active stiffness

$$k_c(t) = k_0 \int_{\mathbb{R}} n(s,t)\, ds$$

• Control signal

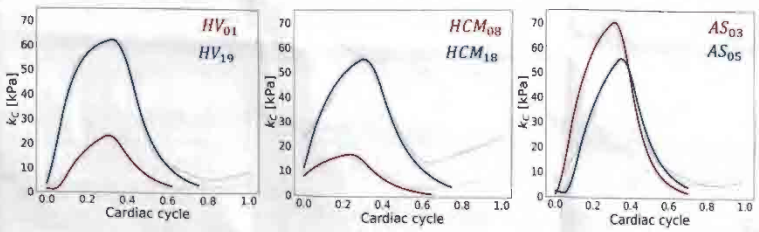
$$u(t) = \begin{cases} k_{\mathrm{ATP}}, & t \in [T_d,\ T_d + \mathrm{APD}], \\ -k_{\mathrm{RS}}, & t \in [T_d + \mathrm{APD},\ T_r], \\ 0, & \text{otherwise} \end{cases}$$

• Active stiffness rate

$$\dot{k}_c = \big(u(t) + \alpha|\dot{e}_c|\big)k_c + n_0 k_0 k_{\mathrm{ATP}} \mathbb{1}_{\{[Ca^{2+}]>C\}}$$

Results

- **20** healthy volunteers (**HV**) • **20** patients with hypertrophic cardiomyopathy (**HCM**) • **5** patients with aortic stenosis (**AS**)



Group	ID	k_0 [kPa]	k_{ATP} [s^{-1}]	k_{SR} [s^{-1}]
HV	HV_01	23.60	20.86	13.47
	HV_19	62.81	25.96	10.00
HCM	HCM_08	19.33	10.07	7.59
	HCM_18	61.34	10.67	7.28
AS	AS_03	78.77	10.46	10.02
	AS_05	69.09	7.22	7.28

Parameter	HV	HCM	AS
n (subjects)	20	20	5
MAE [kPa]	2.29 ± 1.05	8.69 ± 12.58	3.35 ± 1.74
k_0 [kPa]	48.79 ± 16.49	52.22 ± 29.43	$72.19 \pm 5.04^{**}$
k_{ATP} [s^{-1}]	17.66 ± 4.99	$16.46 \pm 9.79^{*}$	12.04 ± 4.42
k_{SR} [s^{-1}]	11.63 ± 1.93	$7.35 \pm 0.11^{*}$	8.99 ± 1.91

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, versus HV, Mann–Whitney U test.

Main takeaways

CLINICAL GAP
MS is clinically relevant but not routinely quantified; existing tools are indirect or invasive.

FAST SWE-DRIVEN SIMULATIONS
SWE-informed 0D model enables fast, interpretable initialization of 3D simulations.

BRIDGING IMAGING & MECHANICS
SWE has the potential to move from descriptive imaging to a quantitative driver of predictive simulation

PATH-READY PIPELINE
k_{ATP} (XB cycling) and k_{SR} (Ca^{2+} re-uptake) ensuring physiological interpretability

Building on **our previous work [1]**:
Myocardial Stiffness Quantification using Ultrasound Shear Wave Elastography and Reduced Modeling for Subject-Specific Simulations, C. Ferrario, JR. Padilla, M. Venet, O. Villemain, M. Sermesant, International Conference on Functional Imaging and Modeling of the Heart, 41-54, Springer 2025.

[2] Villalobos Lizardi, et al., 2022 - Nat. Cardiovasc. Res.
[3] Villemain et al., 2020 - JACC Cardiovasc. Imaging
[4] Bestel et al., 2001 - MICCAI
[5] Caruel et al., 2014 – Biomech. Model. Mechanobiol.
[6] Chapelle et al., 2012 - Int. J. Multiscale Comput. Eng.
[7] Hansen, N.: *The CMA Evolution Strategy: A Tutorial*, 2016
[8] Mollero et al., 2018 - Med. Image Anal.

Bioinformatics Workflows: Representation and Query

ShareFAIR

Mouna El Garb, Emmanuel Coquery, Fabien Duchateau, Nicolas Lumineau

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Context

ShareFAIR project aims to develop a biomedical data analysis platform for the sharing and reuse of scientific **workflows**, with a query engine designed to address the **findable** aspect of FAIRness.

Scenario

From the user's query to relevant workflows.

Problematic

- How to sufficiently **describe** scientific workflows for reuse and sharing ?
- How to write and build **queries** to **retrieve** relevant workflows ?

State of the art

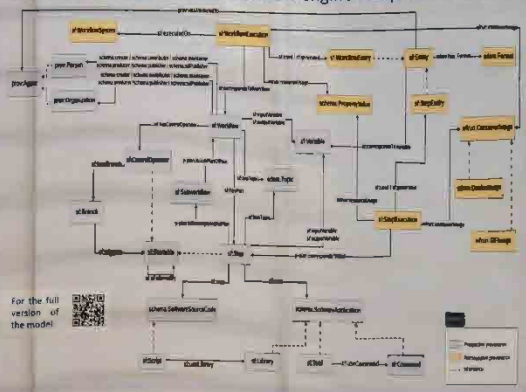
Overview of the features of the BioFlow-Model, compared to the CWL and P-Plan models, and WRROC profiles of RO-Crate.

Category	Subcategory	CWL	P-Plan	WRROC	BioFlow-Model
Scientific context	Workflow design	✔	~	✔	✔
	Entity annotations	~	~	✔	✔
	Workflow execution aim		~	✔	✔
Data	Data identification	~	~	✔	✔
	File characteristics	~		✔	~
	Data access	~		~	✔
	Parameter mapping	✔	~	✔	✔
Software	Software identification	✔		✔	✔
	Software documentation	✔		✔	✔
	Software access	✔		✔	✔
Workflow	Workflow software	✔	-	✔	✔
	Workflow parameters	✔	~	✔	✔
	Workflow requirements	✔	-	~	-
Computational environment	Software environment	-			-
	Hardware environment	-			-
	Container image	~		~	~
Execution detail	Execution timestamps		~	✔	✔
	Consumed resources			✔	✔
	Workflow engine			✔	✔
	Human agent		✔	✔	✔
Control flow			-		✔

Legend: ✓ fully represented, ~ partially represented, - not covered

BioFlow-Model

A structured model to describe workflows using metadata and provenance in a generic and workflow engine independent format.



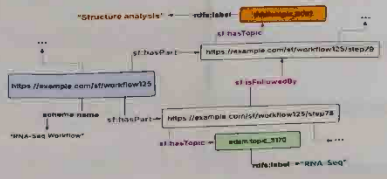
For the full version of the model

Example of query with BioFlow-Model

Query: Find the workflows having an **RNA-Seq** step followed by a **Sequence analysis** step.

```
FIND wf
FROM Workflow wf, Step step1, Step step2
WHERE step1 IN wf
AND step2 IN wf
AND step1.topic XLIKE "RNA-Seq"
AND step2.topic XLIKE "Structure analysis"
AND step1 ISJUSTBEFORE step2;
```

- Graph of the workflow matching to the query:



Perspectives

- Experimental validation of BioFlow-Model
- Rewrite the queries to extract workflows that closely match the user's query
- Rank the workflows returned by query execution based on their relevance

Mouna EL GARB

Théotime FEHR--DELUDE

Unsupervised Anomaly Detection of MRI brain lesions using denoising diffusion probabilistic models

Théotime Fehr Delude[1], Loïc Legris[1,2], Benjamin Lamasson[1], Emmanuel Barbier[1]
[1]Univ. Grenoble Alpes, Inserm U1216, Grenoble Institut des Neurosciences, Grenoble, France,
[2]CHU Grenoble Alpes, Univ. Grenoble Alpes, 38000 Grenoble, France

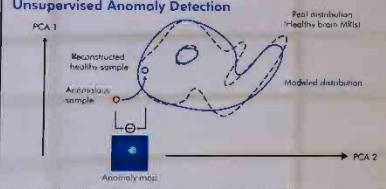
brain health trajectories

GIN

Context

The ability to better quantify the Brain Health Status could lead to a more precise prediction of its Brain Health Trajectory. Unsupervised segmentation of brain lesions leverages large datasets of healthy brain MRI images to detect novel anomalies without the need for training for specific lesions using manual ground truth labeling

Unsupervised Anomaly Detection



At train time, the distribution of a healthy brain MRI dataset is learned. At test time, an anomalous brain MRI is reconstructed to its healthy equivalent. The difference between the two gives an anomaly map which can reveal lesions.

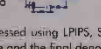
Material & Methods

Denoising Diffusion Probabilistic Models [1]

Training

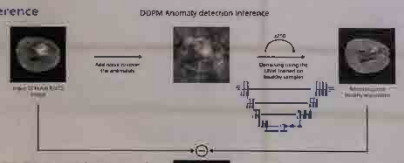
Two models were developed: a T2 FLAIR MRI model trained on 1331 healthy brain MRIs obtained from the OASIS, The Dallas Lifespan Brain Study and The Max Planck Institute LEMON datasets and an ADC model trained on 1952 healthy brain MRIs from The Dallas Lifespan Brain Study, the Human Connectome project and the IXI datasets. Both models were trained on the 72 2D middle axial slices of every 128x128x128 MRI volumes registered to an MNI template and harmonized by intensity.

At train time, denoising diffusion probabilistic models successively add noise to an input image and a UNet-based architecture learns to denoise the image. The denoising capability of each model was assessed using LPIPS, SSIM and PSNR image similarity metrics between the original image and the final denoised image.



Simplex noise was used for its low frequency components that better cover anomalies at test time with less needed noise [2].

Inference

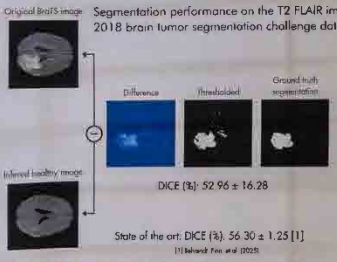


The T2 FLAIR model was tested on the BraTS 2018 brain tumor segmentation dataset and on the ISLES-2022 ischemic stroke lesion segmentation dataset. The ADC model was tested on the ISLES-2022 datasets. At inference time noise is added to the image to cover anomalies, then the UNet denoises the image to reconstruct a healthy equivalent of the input image. The difference gives an anomaly map.

A final 3D anomaly map is obtained by stacking the 2D infered slices which are the thresholded to obtain a binary mask. The noise level and segmentation threshold were chosen for the best performance on a separate subdivision of the test dataset. The final segmentation performance was evaluated using the DICE score.

Experiments & Results

BraTS 2018 brain tumor segmentation



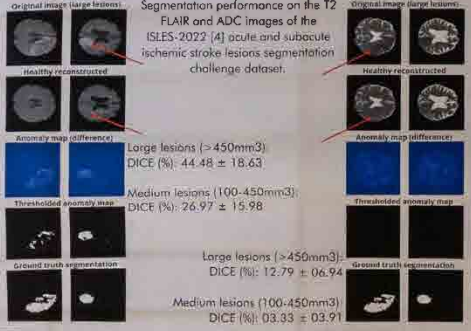
Segmentation performance on the T2 FLAIR images of the BraTS [3] 2018 brain tumor segmentation challenge dataset

DICE (%): 52.96 ± 16.28

State of the art: DICE (%): 56.30 ± 1.25 [1]

[1] Behrendt Fabian et al. (2023)

ISLES-2022 ischemic stroke lesion segmentation



Segmentation performance on the T2 FLAIR and ADC images of the ISLES-2022 [4] acute and subacute ischemic stroke lesions segmentation challenge dataset.

Large lesions (>450mm3):
DICE (%): 44.48 ± 18.63

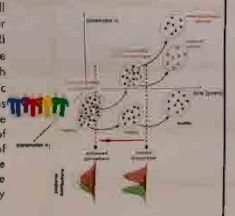
Medium lesions (100-450mm3):
DICE (%): 26.97 ± 15.98

Large lesions (>450mm3):
DICE (%): 12.79 ± 06.94

Medium lesions (100-450mm3):
DICE (%): 03.33 ± 03.91

Future works & Conclusion

The anomaly detection framework will be refined to work better on smaller lesions, will be extended to other MRI contrasts (T1w and DWI) and will be challenged on other types of lesions with the aim of leveraging multiparametric acquisitions for multiple anomaly maps per patient. The end goal is to create tools that encompass a large variety of lesional and non-lesional markers of brain health status which can then be used on individual brains to detect the ones which will deviate from a healthy trajectory.



The computations were performed using the GRICAD infrastructure (https://gricad.univ-grenoble-alpes.fr)

GIN · CHU Grenoble Alpes · brain health trajectories · France 2030 Programme d'Investissements d'Avenir

References

(1) Ho, Jonathan, et al. "Denoising diffusion probabilistic models." (2020).
(2) Wyatt, Julian, et al. "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise." (2022).
(3) Menze, Bjoern H., et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." (2014).
(4) H. Petzsche, Moritz R., et al. "ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset." (2022).

Laurène ASSAILLY

Laurène Assailly - Projet SaNSo
laurene.assailly@ens.psl.eu
Institut Santé Numérique en Société
ENS, Centre Maurice Halbwachs

APPRENTISSAGE FÉDÉRÉ, DES ENJEUX POLITIQUES
CAS DU GOUVERNEMENT DES DONNÉES HOSPITALIÈRES

La **réutilisation scientifique de données issues des systèmes de santé** est une priorité des politiques nationales de recherche et de stratégies d'innovation de l'Union Européenne. Dans ce cadre, l'**apprentissage fédéré** est mis en avant comme une technique protectrice de la confidentialité des données. C'est aussi un **instrument politique** pour certains acteurs de la recherche. Des **réseaux de recherche hospitaliers** qui promeuvent la réutilisation de ces données sont pourtant **réticents à participer aux infrastructures** dédiées et **développent des alternatives**. Ils investissent l'apprentissage fédéré comme un moyen de ne pas recourir à des plateformes nationales.

Pourquoi cette défiance envers les politiques de réutilisation des données de santé ?
En quoi investir l'apprentissage fédéré est une position à la fois scientifique et politique ?

Méthodes : enquête qualitative en science politique. 60 entretiens semi-directifs, 1 an et demi d'observations ethnographiques.

Contrôler les données hospitalières, levier et enjeu du travail médical

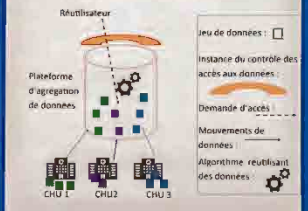
Produire des données de santé et les traduire en données de recherche est un **travail** souvent **invisible** (Denis, 2018) qui nécessite une **expertise**. Dans les hôpitaux, des médecins spécialisé·es en informatique médicale prennent en charge ces tâches, qui sont peu reconnues (Assailly & Thomas, 2024). Ils et elles n'ont pas d'activité clinique. Prendre soin des données des patient·es et assurer le secret médical à l'hôpital structure **leur identité professionnelle**.

Ces médecins volent donc les **politiques publiques** qui visent à faire circuler les données vers des plateformes nationales comme une menace. En **détachant les données de l'hôpital**, cette infrastructure peut les déposséder du produit de leur travail.

Ces **données** sont cependant **spécifiques à l'infrastructure** hospitalière où **elles sont créées**. Les réutiliser en recherche nécessite la participation des spécialistes locaux des données, afin de les rendre intelligibles pour d'autres usages. Par leur expertise située, les médecins responsables des données hospitalières disposent d'une **capacité d'action sociotechnique** (Mitchell, 2017) : **sans leur travail de traduction, accéder aux données ne suffit pas à les comprendre ou à les réutiliser**.

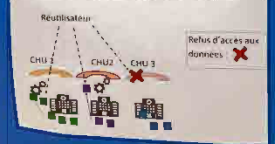
Infrastructure d'agrégation de données

Gouvernement centralisé des accès aux données



Infrastructure fédérée

Sélection ou refus local des données mises en accès



Investir l'apprentissage fédéré : un exemple de la dimension politique des outils et des infrastructures

La valorisation – ici scientifique – **de données** nécessite un travail. Cela crée des **tensions** quand les personnes qui les produisent ne sont pas celles qui en tirent bénéfice. Or, le partage de la valorisation dépend de l'infrastructure sociotechnique en place. Les outils et infrastructures informatiques sont des **instruments de gouvernement** (Lascoumes & Le Galès, 2005) : ils participent à définir la part des différents acteurs dans le travail scientifique et ses rétributions (reconnaissance, citations...)

Les acteurs hospitaliers responsables du travail de production de données **s'opposent** aux infrastructures qui déplacent le **contrôle des réutilisations des données au niveau national**. De tels instruments de gouvernement des données réduisent la maîtrise de la **reconnaissance scientifique et des rétributions** auxquelles aspirent les acteurs qui produisent les données de recherche lors de réutilisations.

Pour que la recherche sur données de santé se développe et produise tout de même des savoirs en santé, d'**autres types d'infrastructures locales** sont construites. Elles établissent un **autre équilibre des pouvoirs** dans le gouvernement des données.

- Les responsables du **réseau de recherche hospitalier** s'intéressent aux méthodes d'**apprentissage fédéré**. Faire circuler les algorithmes d'un hôpital à l'autre plutôt que de déplacer les données vers une base centralisée pour les réutiliser **permet que les données** restent sous **contrôle des institutions qui les produisent**. Pas de risque de détachement des données : elle ne circulent pas hors des hôpitaux.
- Pour les travaux qui nécessitent de disposer des données de plusieurs hôpitaux dans un même espace, plutôt que de recourir à des plateformes tierces, construire un ***data center***. Les données sont **stockées localement** dans cette infrastructure. Les **accès aux données** ont lieu dans cet espace informatique contrôlé. Seuls les résultats des analyses en sont extraits : pas les données, elles sont **réutilisées sans circuler**.

Ces infrastructures sont construites par des médecins, **au service de leur groupe professionnel** et de ses valeurs :

- disposer d'espaces autonomes **de production des savoirs médicaux** ;
- assurer la **protection du secret médical** sous contrôle médical ;
- **valoriser** le travail des données et **l'expertise des hôpitaux publics**.

Choisir l'apprentissage fédéré : un positionnement implicite dans des débats politiques

Les réutilisations de données de santé ne sont pas contestées par les médecins technophiles spécialisé·es en informatique médicale. Ces acteurs s'opposent au **rapport de force que créent des plateformes de centralisation des données**.

Construire des infrastructures et des méthodes d'apprentissage fédéré parmi d'autres choix techniques - leur permet donc de renforcer leur position politique et scientifique en contrôlant les données. On voit ici la **dimension intrinsèquement politique des techniques**. Les choix d'infrastructures sont connectés à des questions politiques pour définir les réutilisations désirables des données :

- **Que sont les données de santé** : un outil de soin et de savoir médical ? Ou bien une ressource valorisable par des acteurs étrangers à leur production ?
- **Quelles sont les priorités des réutilisations de données** : produire des savoirs déterminés par les médecins, pour les acteurs du système de santé ? Ou bien des projets d'innovation dirigés par des acteurs extérieurs aux activités de soin ?
- Concernant le **périmètre de la profession médicale** : le travail des données doit-il en faire partie ? Dans quelle mesure la production de savoirs médicaux doit-elle être orientée par des médecins ?

Bibliographie :

Assailly, L., & Thomas, F. (2024). Stratégies médicales de mise en (in)visibilité des médecins DIM. *Revue des sciences sociales*, *72*.

Mitchell, T. (2017). *Carbon democracy. Le pouvoir politique à l'ère du pétrole*. la Découverte.

Lascoumes, P., & Le Galès, P. (Eds.). (2005). *Gouverner par les instruments*. Presses de Sciences Po.

Denis, J. (2018). *Le travail invisible des données. Éléments pour une sociologie des infrastructures scripturales*. Presse des Mines.

En savoir plus :

Assailly, L. (2024). *La santé des données. Enquête sur un travail hospitalier de production, de contrôle et de mise en accès de données de santé face aux stratégies d'action publique pour les valoriser*. Thèse de doctorat, Université de Strasbourg. https://hal.science/tel-04984388

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0004
Travaux réalisés dans le cadre d'une recherche doctorale financée par la Région Grand Est

Federated querying of genomic health data leveraging semantic web technologies and the Beacon standard for genomic discoverability

Alexandrina Bodrug-Schepers [1], Hugo Chabane [2], Gabriela Montoya [2], Patricia Serrano-Alvarado [2], Richard Redon [1], Alban Gaignard [1,3]

1. Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
2. Nantes Université, LS2N, Nantes, France
3. IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France

l'institut du thorax
umr1087.univ-nantes.fr

Alexandrina
BODRUG

SCIENTIFIC BACKGROUND

Biomedical challenge

Intracranial aneurysms (ICA) are neurovascular disorders characterised by saccular dilatations of cerebral vessels. They occur in 3 to 6% of the general population. Rupture is the only complication of ICA and although uncommon i.e. in less than 1% of carriers, it leads to death or severe disability in 2/3rds of rupture cases. Understanding ICA formation and rupture is challenging because it is a multifactorial complex disease with few screening solutions.

Project context

The ICAN project [1], a French non-interventional multicenter nationwide study, recruited individuals with ICA and collected neuroimaging, genetic, and phenotypic data. One of the objectives of the **PEPR Santé Numérique Neurovasc** is to make the ICAN dataset more aligned with **FAIR principles** (Findable, Accessible, Interoperable, Reusable) [2].

FAIRification STRATEGIES

Existing solutions

Organisations such as **GA4GH**, **ELIXIR**, and **F-EGA** are collectively developing technical standards, policies, and software to facilitate genomic biomedical research. For the **FAIRification of the ICAN dataset**, we explored two main approaches.

The **Beacon standard** [3] focuses on genomic data discoverability and offers a query framework that incorporate **privacy** considerations. The Beacon standard is also a **REST API** specification.

Semantic web technologies make biomedical concepts and their relationships explicit through specialized ontologies, semantic data models, and **knowledge graphs (KGs)**. Many specialised high quality life science KGs are open access.

FAIR technological demonstrators

We developped **FAIR genomic technological demonstrators** to showcase different stategies, their advantages and their shortcomings.

A FULL SEMANTIC APPROACH

We investigated broadly used open source domain expert ontologies. We chose the **Swiss Personalized Health Network** (SPHN) [4] framework to semantify our pheno-clinical data following their data model. We added descriptions using the **Human Phenotype Ontology** (HPO) [5]. We used several ontologies to represent genomic variation (SO, GENO) and genomic location (FALDO) [6], and developped our **own genomic variation data model** tailored to our needs.

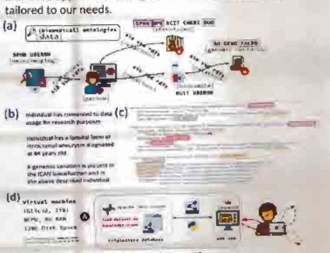


FIGURE 3 : Integrating genomic health data as a KG
Existing biomedical ontologies (a) can be used to add semantics to health genomic data. This approach permits the precise representation of the dataset's concepts and relationships (b) in a machine readable RDF format (c). The RDF file can then be loaded in a triplestore database (as a KG), allowing automated reasoning and querying of multimodal health data ((d) - pink arrows), and integration with public KGs ((d) - blue and yellow dotted arrows).

Take away message

This approach responds to our 2nd and 3rd challenge. It **semantifies**, **integrates** and makes **multi-modal health data queryable** thanks to the **SPARQL query language**. Identical queries could be **distributed** to several organisation using the same framework *i.e.* Swiss Hospitals and Universities. Moreover, thanks to the **federation features** of SPARQL, the ICAN KG becomes **interoperable with UniprotKB and Wikidata**, making the FIGURE 2 use case question possible.

Alexandrina Bodrug
alexandrina.bodrug@univ-nantes.fr
Unité de Recherche de l'institut du Thorax
[illegible] - 8 quai Moncousu 44007 NANTES

DATASET : ICAN BIOCOLLECTION AND PHENO-CLINICAL DATA

Individuals carrying ICA were recruited for the ICAN research projet and benefited from **neuroimaging, biosampling, genetic screening** and **pheno-clinical data** collection. The ICAN dataset is **heterogenic**, multi-centric and **multimodal**. As a genomic health dataset it has **data privacy** protection and **data governance** constraints.

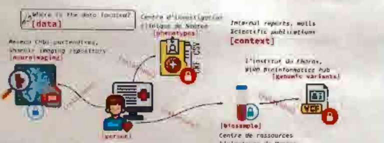


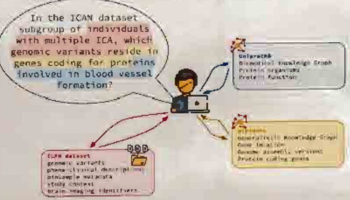
FIGURE 1: THE ICAN DATASET
The ICAN dataset is made of several multimodal domain-specific datasets. These are located in specialized laboratories that have technologies tailored to their specific needs. Privacy risk is higher with pheno-clinical data than with brain imaging, so different sharing constrains apply. Relationships between data chunks are often hard to retrieve and sparsely documented. Context of the study describing for example inclusion and exclusion criteria, can be found in scientific publications, internal reports and mails.

We simulated a synthetic dataset to be used for our publicly available **FAIR technological demonstrators**. We simulated phenotypic data by mimicking per column variable problabilities and eliminating multually exclusive variables (i.e. no gestational diabetes in males). Genetic data was simulated using the methods described in the poster ***« Incorporating complex genetic model into risk stratification »*** (Laporte et al.)

INFO POINT : SYNTHETIC DATASET

CHALLENGES AND MOTIVATION USE CASE SCENARIO

We identified three main challenges to address : (1) **Share sensitive data** in a safe way , (2) **Semantify, integrate** and **query multimodal** health data and (3) **Integrate** health data with **public knowledge bases**.
We constructed a motivating question designed to showcase the technological solutions addressing our challenges and enabling the **FAIRification of the ICAN dataset**.



KGs structure complex concepts and their relationships. They can be generalistic or more specialized, as is the case with **biomedical KGs**. Some example include the biomedical **UniprotKB** focusing on protein knowledge, and the generalistic cross-domain open KG **Wikidata**. All KGs are not open access. Their goals can include acting as knowledge integration hubs, FAIR sharing, automated reasoning and inference aid, or as a domain specific decision support.

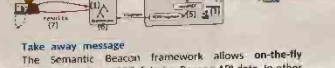
INFO POINT: BIOMEDICAL KGs

FIGURE 2: MOTIVATION USE CASE SCENARIO
Answering our motivational question requires drawing on knowledge from several separate sources. We consider the ICAN dataset as a single integrated source built to be interoperable with existing biomedical KGs through common controlled vocabularies and data models.

SEMANTIC BEACON FRAMEWORK

We first deployed a Beacon API following an existing implementation [7]. This required data transformation to make the ICAN dataset follow the Beacon standard data model. A Beacon API deployer can configure the API **response granularity** : **Boolean** responses, **Aggregated** responses or **Record level** responses. This allows making the data discoverable by signaling its existence (**Boolean**) or sharing statistics (**Aggregated**), without giving detailed (**Record level**) access to unauthorized requests. This is a **powerful feature for sensitive data sharing** and adresses our 1st challenge. We then developped the **Semantic Beacon framework** [8] to make the communication possible betwen the Beacon REST API and existing biomedical knowledge graphs.

FIGURE 4: Semantic Beacon framework
The Semantic Beacon framework uses an ODMTP server as a mediator between the REST API and KGs. On Demand Mapping of Triple Patterns (ODMTP) framework integrates non-RDF datasets on-demand into Linked Data using the Triple Pattern Fragment (TPF) methodology [9]. Query decomposition into tripple lattern queries (TPQ) is delegated to the client (1). These are then passed to the TPF server (2) that uses a Semantic RML Mapping (3) of the genomic variants in the Beacon API to convert the TPQ into a HTTP requests (4). The API response is then mapped back to fragments (5) following the same RML mapping and the client performs the final reconstruction (6) (Execution plan) before sending back a Linked Data response (7).



Take away message

The Semantic Beacon framework allows **on-the-fly integration of non-RDF data** i.e. Beacon API data. In other words genomic variants found within our Beacon API can be enriched with **fresh** public knowledge from KGs **without data duplication** or manual intergration. This framework responds to our 3rd challenge of **integration with public KGs** and offers an **innovate combination** of the biomedical research approach and the semantic web approach for sharing genomic health data knowledge. The Semantic Beacon was successfully deployed as a Proof Of Concept [8], however it is **still experimental** and has some limitations such **slow query execution**.

USE CASE SPARQL QUERY

SPARQL querying allows knowledge retrieval from KGs and from the Semantic Beacon. The question from motivating use case scenario in FIGURE 2 can translated to such a query.

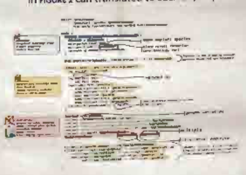


FIGURE 5: Federated SPARQL querying
The query fetches information about proteins and their function in UniportKB, about gene coding products and gene location in Wikidata and about genomic variation location and their association to phenotypes in the ICAN KG or the Semantic Beacon API. The FALDO ontology (in green) is used to precisely describe variation location. The query looks for the same graph structure as the ones represented in the RDF file in FIGURE 3-(c)

REFERENCES

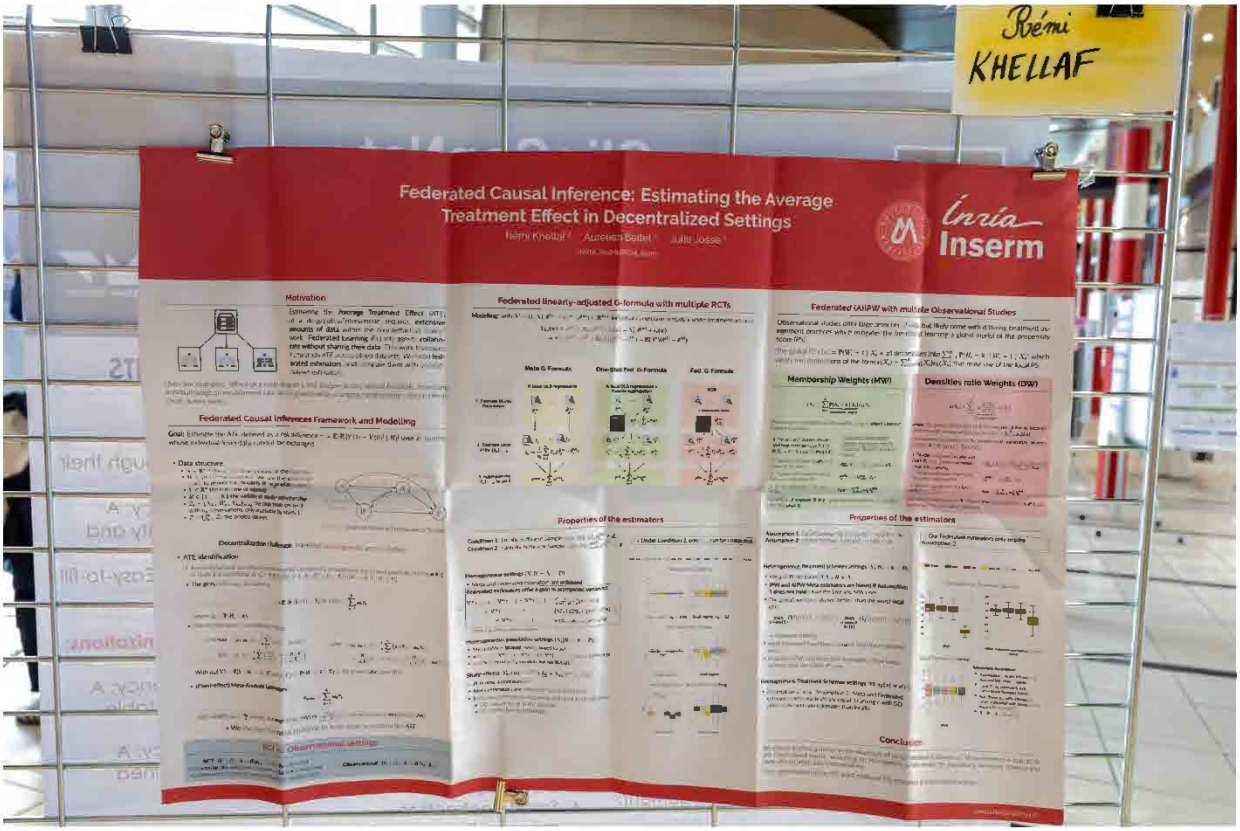
[illegible]

ACKNOWLEDGMENTS

[illegible]

Funded by ANR 22-PESN-00[illegible]

GLiCID Nantes Université Inserm CNRS CHU NANTES


Rémi
KHELLAF
Federated Causal Inference: Estimating the Average Treatment Effect in Decentralized Settings
Inria
Inserm
Motivation
Federated Causal Inferences Framework and Modelling
Federated linearly-adjusted G-formula with multiple RCTs
Federated IAIPW with multiple Observational Studies
Membership Weights (MW)
Densities ratio Weights (DW)
Properties of the estimators
Properties of the estimators
Conclusion

Montassar NAGHMOUCHI

ClinConNet

Privacy preserving identity & consent management for clinical trials

TELECOM SudParis — IP PARIS — Institut Mines-Télécom

Authors

Montassar Naghmouchi
Maryline Laurent

samovar

Partners

LATIM — LTSI — LASCO — cea list — CHU Rennes — LIFO Laboratoire d'Informatique Fondamentale d'Orléans — anr — France 2030

1.THE CHALLENGE

Complex ond outdated consent management.
→Paper-based, slow and inefficient.
→Patient data locked in siloed systems.
→Limited patient control and involvement.

2. THE SOLUTION

A modern patient-centric approach combining Blockchain, SSI and Dynamic consent.

Web-Portal: Easy-to-use for project discovery and consent management.

SSI wallet: Patient-centric identity and consent management.

Blockchain: Secure, immutable record of consent.

3. HOW IT WORKS

1. SSI wallets to manage identity across multiple projects.
2. Consent forms exchanged as wallet-to-wallet messages.
3. Consent forms encoded as HTML, signed by participants and research organizations.
4. Consent proofs generated from consent forms.
5. Public consent proofs available on the immutable blockchain.
6. Smart contracts to manage the consent state.

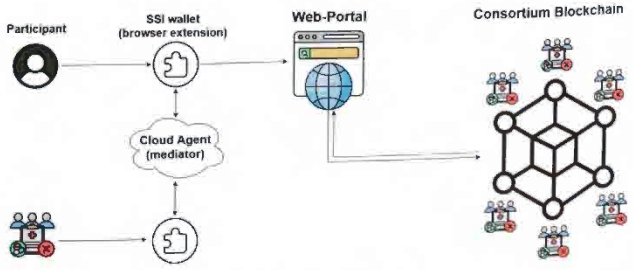


Figure 1 ClinConNet Overview

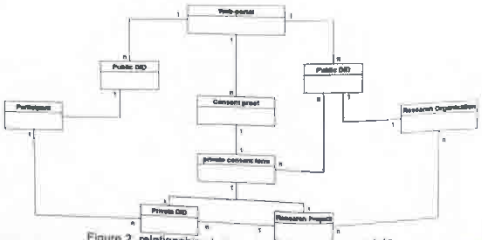


Figure 2 relationships between actors and consent data

4. KEY BENEFITS

For **Participants**:

- Full Control: Through their wallet.
- Complete Privacy: A protected identity and consent data.
- Simple & Clear: Easy-to-fill digital forms.

For **Research Organizations**:

- Trust & Transparency: A secure and auditable process.
- Greater Efficiency: A modern, streamlined workflow.

TRACIA

Contact montassar-bellah_naghmouchi@telecom-sudparis.eu
maryline.laurent@telecom-sudparis.eu

Secure, Safe and Fair Federated Learning for Healthcare

Hichem FARAOUN

LATIM Laboratoire de Traitement de l'Information Médicale

Inserm

Institut Mines-Télécom

Authors

Hichem Faraoun,
PhD Student, IMT Atlantique

Reda Bellafqira
Professor, IMT Atlantique

Gouenou Coatrieux
Professor, IMT Atlantique

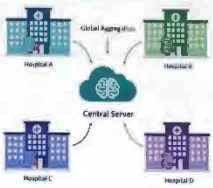
Kassem Kallas
Professor, IMT Atlantique

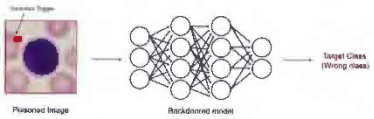
Project

Projet financé dans le cadre de France 2030 par l'ANR sous la référence PEPR-SN SSF-ML-DH

Motivation:

Federated Learning (FL) enables collaborative model training across institutions without sharing raw patient data, making it well-suited for healthcare. However, FL is vulnerable to adversarial threats, particularly backdoor attacks, which pose a serious risk to medical applications.





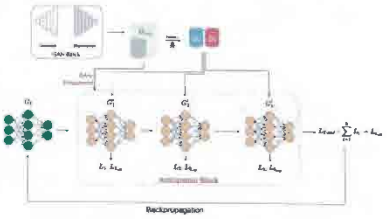
What is a backdoor attack?

A backdoor attack compromises training by inserting a small trigger into some training samples and relabeling them to a chosen target class. The model learns to associate the trigger with this target, so at inference time, any input containing the trigger is misclassified, while performance on clean inputs remains unchanged.

Proposed Method:

We propose a two-stage backdoor attack, consisting of the following steps:

1. GAN-based data alignment: we train a GAN guided by the global model to synthesize samples from the global distribution, reducing the distribution gap caused by non-IID data.
2. Anticipation of future updates: We simulate benign clients' updates and anticipate future global model changes, to optimize the malicious update accordingly



Threat Model:

The attacker controls up to 10% of clients, and has access to their local data, and local model training, but does not know or control other clients' private data nor the server's aggregation rule.

Results:

The attacker has 20% of chances of injecting the backdoor during the first training window [0-100]. Our method achieves substantially higher persistence, across all three scenarios on different MedMNIST datasets, highlighting its effectiveness in realistic FL scenarios.

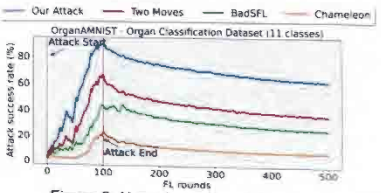


Figure 2: Norm-bounding defense

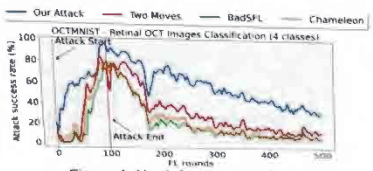


Figure 1: No defense scenario

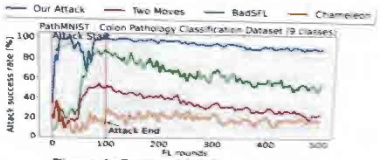


Figure 3: Foolsgold defense

Conclusion & Future Work:

Our results reveal the susceptibility of federated learning to persistent backdoor attacks, highlighting potential risks for medical AI systems. In future work, we will focus on designing and evaluating healthcare-specific defense mechanisms to ensure model safety and reliability.

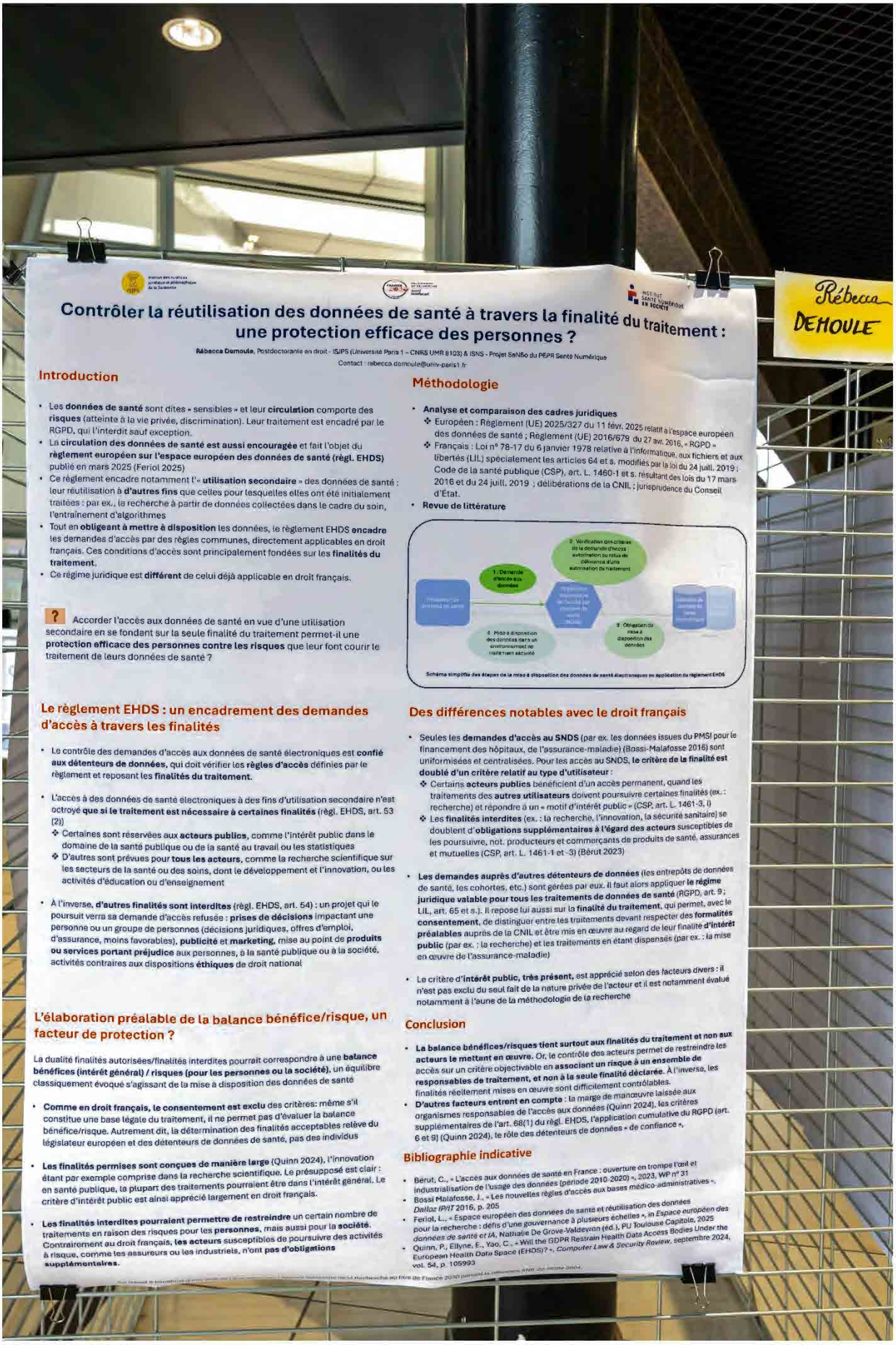
References

[1] McMahanB, et al. Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. PMLR, 2017.

[2] Wen Y, Geiping J, Fowl L, Souri H, Chellappa R, Goldblum M, et al. Thinking two moves ahead: Anticipating other users improves backdoor attacks in federated learning. arXiv preprint arXiv:2210.09305. 2022.

[3] Dai Y, Li S. Chameleon: Adapting to peer images for planting durable backdoors in federated learning. In: International Conference on Machine Learning. PMLR; 2023.

[4] Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V. How to backdoor federated learning. In: International conference on artificial intelligence and statistics. PMLR; 2020.



Contrôler la réutilisation des données de santé à travers la finalité du traitement : une protection efficace des personnes ?

Rébecca Demoule, Postdoctorante en droit - ISJPS (Université Paris 1 – CNRS UMR 8103) & ISNS - Projet SoNBo du PEPR Santé Numérique
Contact : rebecca.demoule@univ-paris1.fr

Introduction

- Les **données de santé** sont dites « sensibles » et leur **circulation** comporte des **risques** (atteinte à la vie privée, discrimination). Leur traitement est encadré par le RGPD, qui l'interdit sauf exception.
- La **circulation des données de santé est aussi encouragée** et fait l'objet du **règlement européen sur l'espace européen des données de santé (règl. EHDS)** publié en mars 2025 (Feriol 2025)
- Ce règlement encadre notamment l'« **utilisation secondaire** » des données de santé : leur réutilisation à **d'autres fins** que celles pour lesquelles elles ont été initialement traitées : par ex., la recherche à partir de données collectées dans le cadre du soin, l'entraînement d'algorithmes
- Tout en **obligeant à mettre à disposition** les données, le règlement EHDS **encadre** les demandes d'accès par des règles communes, directement applicables en droit français. Ces conditions d'accès sont principalement fondées sur les **finalités du traitement.**
- Ce régime juridique est **différent** de celui déjà applicable en droit français.

? Accorder l'accès aux données de santé en vue d'une utilisation secondaire en se fondant sur la seule finalité du traitement permet-il une **protection efficace des personnes contre les risques** que leur font courir le traitement de leurs données de santé ?

Méthodologie

- **Analyse et comparaison des cadres juridiques**
  - ❖ Européen : Règlement (UE) 2025/327 du 11 févr. 2025 relatif à l'espace européen des données de santé ; Règlement (UE) 2016/679 du 27 avr. 2016, « RGPD »
  - ❖ Français : Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (LIL) spécialement les articles 64 et s. modifiés par la loi du 24 juill. 2019 ; 2016 et du 24 juill. 2019 ; délibérations de la CNIL ; jurisprudence du Conseil d'État.
- **Revue de littérature**

Schéma simplifié des étapes de la mise à disposition des données de santé électroniques en application du règlement EHDS

Le règlement EHDS : un encadrement des demandes d'accès à travers les finalités

- Le contrôle des demandes d'accès aux données de santé électroniques est **confié aux détenteurs de données,** qui doit vérifier les **règles d'accès** définies par le règlement et reposant les **finalités du traitement.**
- L'accès à des données de santé électroniques à des fins d'utilisation secondaire n'est octroyé **que si le traitement est nécessaire à certaines finalités** (règl. EHDS, art. 53 (2))
  - ❖ Certaines sont réservées aux **acteurs publics**, comme l'intérêt public dans le domaine de la santé publique ou de la santé au travail ou les statistiques
  - ❖ D'autres sont prévues pour **tous les acteurs**, comme la recherche scientifique sur les secteurs de la santé ou des soins, dont le développement et l'innovation, ou les activités d'éducation ou d'enseignement
- À l'inverse, **d'autres finalités sont interdites** (règl. EHDS, art. 54) : un projet qui le poursuit verra sa demande d'accès refusée : **prises de décisions** impactant une personne ou un groupe de personnes (décisions juridiques, offres d'emploi, d'assurance, moins favorables), **publicité** et **marketing**, mise au point de **produits ou services portant préjudice** aux personnes, à la santé publique ou à la société, activités contraires aux dispositions **éthiques** de droit national

Des différences notables avec le droit français

- Seules les **demandes d'accès au SNDS** (par ex. les données issues du PMSI pour le financement des hôpitaux, de l'assurance-maladie) (Bossi-Malafosse 2016) sont uniformisées et centralisées. Pour les accès au SNDS, **le critère de la finalité est doublé d'un critère relatif au type d'utilisateur :**
  - ❖ Certains **acteurs publics** bénéficient d'un accès permanent, quand les traitements des **autres utilisateurs** doivent poursuivre certaines finalités (ex. : recherche) et répondre à un « motif d'intérêt public » (CSP, art. L. 1461-3, I)
  - ❖ Les **finalités interdites** (ex. : la recherche, l'innovation, la sécurité sanitaire) se doublent d'**obligations supplémentaires à l'égard des acteurs** susceptibles de les poursuivre, not. producteurs et commerçants de produits de santé, assurances et mutuelles (CSP, art. L. 1461-1 et -3) (Bérut 2023)
- **Les demandes auprès d'autres détenteurs de données** (les entrepôts de données de santé, les cohortes, etc.) sont gérées par eux. Il faut alors appliquer **le régime juridique valable pour tous les traitements de données de santé** (RGPD, art. 9 ; LIL, art. 65 et s.). Il repose lui aussi sur la **finalité du traitement**, qui permet, avec le **consentement**, de distinguer entre les traitements devant respecter des **formalités préalables** auprès de la CNIL et être mis en œuvre au regard de leur finalité **d'intérêt public** (par ex. : la recherche) et les traitements en étant dispensés (par ex. : la mise en œuvre de l'assurance-maladie)
- Le critère d'**intérêt public, très présent,** est apprécié selon des facteurs divers : il n'est pas exclu du seul fait de la nature privée de l'acteur et il est notamment évalué notamment à l'aune de la méthodologie de la recherche

L'élaboration préalable de la balance bénéfice/risque, un facteur de protection ?

La dualité finalités autorisées/finalités interdites pourrait correspondre à une **balance bénéfices (intérêt général) / risques (pour les personnes ou la société)**, un équilibre classiquement évoqué s'agissant de la mise à disposition des données de santé

- **Comme en droit français, le consentement est exclu** des critères: même s'il constitue une base légale du traitement, il ne permet pas d'évaluer la balance bénéfice/risque. Autrement dit, la détermination des finalités acceptables relève du législateur européen et des détenteurs de données de santé, pas des individus
- **Les finalités permises sont conçues de manière large** (Quinn 2024), l'innovation étant par exemple comprise dans la recherche scientifique. Le présupposé est clair : en santé publique, la plupart des traitements pourraient être dans l'intérêt général. Le critère d'intérêt public est ainsi apprécié largement en droit français.
- **Les finalités interdites pourraient permettre de restreindre** un certain nombre de traitements en raison des risques pour les **personnes**, mais aussi pour la **société**. Contrairement au droit français, **les acteurs** susceptibles de poursuivre des activités à risque, comme les assureurs ou les industriels, n'ont **pas d'obligations supplémentaires.**

Conclusion

- **La balance bénéfices/risques tient surtout aux finalités du traitement et non aux acteurs le mettant en œuvre.** Or, le contrôle des acteurs permet de restreindre les accès sur un critère objectivable en **associant un risque à un ensemble de responsables de traitement, et non à la seule finalité déclarée.** À l'inverse, les finalités réellement mises en œuvre sont difficilement contrôlables.
- **D'autres facteurs entrent en compte** : la marge de manœuvre laissée aux organismes responsables de l'accès aux données (Quinn 2024), les critères supplémentaires de l'art. 68(1) du règl. EHDS, l'application cumulative du RGPD (art. 6 et 9) (Quinn 2024), le rôle des détenteurs de données « de confiance ».

Bibliographie indicative

- Bérut, C., « L'accès aux données de santé en France : ouverture en trompe l'œil et industrialisation de l'usage des données (période 2010-2020) », 2023, WP n° 31
- Bossi Malafosse, J., « Les nouvelles règles d'accès aux bases médico-administratives », *Dalloz IP/IT* 2016, p. 205
- Feriol, L., « Espace européen des données de santé et réutilisation des données pour la recherche : défis d'une gouvernance à plusieurs échelles », in *Espace européen des données de santé et IA*, Nathalie De Grove-Valdeyron (éd.), PU Toulouse Capitole, 2025
- Quinn, P., Ellyne, E., Yao, C., « Will the GDPR Restrain Health Data Access Bodies Under the European Health Data Space (EHDS)? », *Computer Law & Security Review*, septembre 2024, vol. 54, p. 105993

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-22-PESN-0004.

Cryptographic Commitments on Anonymizable Data

Xavier Bultel[1,2], Céline Chevalier[3], Charlène Jojon[1,2], Diandian Liu[1] and Benjamin Nguyen[1,2]

[1] INSA Centre Val de Loire, Université d'Orléans, France [2] Inria-Saclay, France
[3] CRED, Université Panthéon-Assas, Paris II, DIENS, Ecole normale supérieure, PSL University, CNRS, Inria, France

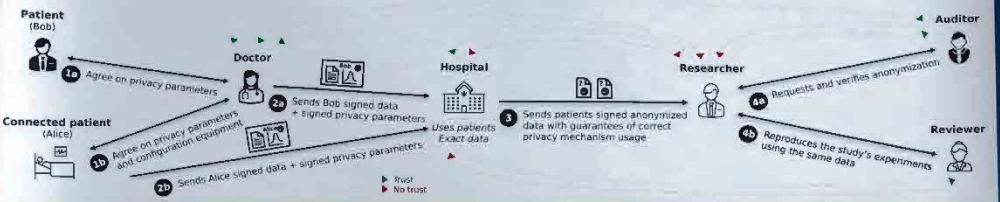
Charlène JOJON

Context and Motivation

Local Differential Privacy (LDP) mechanisms consist of (locally) adding controlled noise to data in order to protect the privacy of their owner.

Objectives of this paper:

- Introduce a new cryptographic primitive called LDP commitment. Usually, a commitment ensures that the committed value cannot be modified before it is revealed. In the case of an LDP commitment, however, the value is revealed after being perturbed by an LDP mechanism. Opening an LDP commitment therefore requires a proof that the mechanism has been correctly applied to the value, to ensure that the value is still usable for statistical purposes.
- Provide a concrete LDP Commitment scheme that achieves an LDP staircase mechanism (Generalized Randomized Response Technique).
- Propose an application to show how our primitive ensure privacy, usability and traceability of medical data when it is used in open science statistical studies.
- Present a security model for the LDP Commitment primitive (Hiding and Binding properties).



Definitions

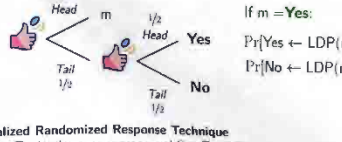
Local Differential Privacy (LDP)

A randomized algorithm $\mathcal{A}$ satisfies *ϵ-local differential privacy* (LDP) where $\epsilon \in \mathbb{R}^{+*}$, if for any pair of input values $(m_1, m_2) \in \text{Domain}(\mathcal{A})$ and any possible output $\hat{m} \in \text{Codomain}(\mathcal{A})$:

$$\Pr[\mathcal{A}(m_1) = \hat{m}] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(m_2) = \hat{m}].$$

Classical Randomized Response Technique

m represents the true response to a binary question, for instance: "Do you have diabetes?"

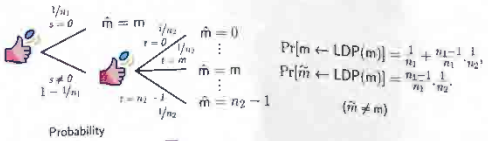


If m = **Yes**:

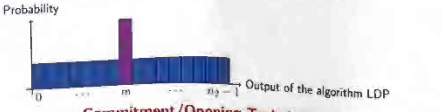
$$\Pr[\text{Yes} \leftarrow \text{LDP}(m)] = \frac{3}{4},$$
$$\Pr[\text{No} \leftarrow \text{LDP}(m)] = \frac{1}{4}.$$

Generalized Randomized Response Technique

Let $\mathcal{M} = \mathbb{Z}_{n_2}$ be the message space, and $\Theta = \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$.
$m \in \mathcal{M}$ denotes the original message, and $\hat{m}$ a random message in $\mathcal{M}$.
$\text{LDP}(m)$: picks $s \in \mathbb{Z}_{n_1}$, if $s = 0$, $\hat{m} = m$, else picks $t \in \mathbb{Z}_{n_2}$, $\hat{m} = t$.



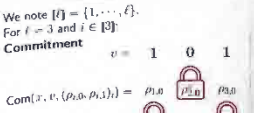
$$\Pr[m \leftarrow \text{LDP}(m)] = \frac{1}{n_1} + \frac{n_1 - 1}{n_1} \cdot \frac{1}{n_2},$$
$$\Pr[\tilde{m} \leftarrow \text{LDP}(m)] = \frac{n_1 - 1}{n_1} \cdot \frac{1}{n_2}.$$
$$(\tilde{m} \neq m)$$



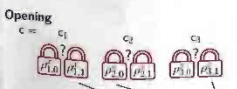
Commitment/Opening Technique

We note $[\ell] = \{1, \cdots, \ell\}$.
For $\ell = 3$ and $i \in [3]$:

Commitment



Opening



Optimized Randomized Response Commitment Scheme (ORRC)

ORRC = (Setup, Commit, VerCommit, Open, VerOpen, OpenLDP, VerOpenLDP).

Computation time (in ms)

(ℓ_1, ℓ_2)	Setup	Commit	VerCommit	Open	VerOpen	OpenLDP	VerOpenLDP
(7, 7)	0.54	6.36	5.72	0.51	0.14	0.85	0.52
(30, 30)	2.25	28.45	25.37	1.71	0.15	1.50	0.56

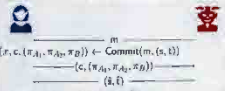
ORRC scheme description

$\text{Setup}(\lambda, \text{LDP}) \to (\lambda, \mathbb{G}, p, g, (g_{j,0}, g_{j,1})_j, (f_{i,0}, f_{i,1}, h_{i,0}, h_{i,1})_i, \mathcal{M}, \Theta)$: $\mathbb{G} = \langle g \rangle$ s.t. $|\mathbb{G}| = p$
For each $j \in [\ell_1]$, picks $g_{j,0}, g_{j,1} \xleftarrow{\$} \mathbb{G}$. For each $i \in [\ell_2]$, picks $f_{i,0}, f_{i,1}, h_{i,0}, h_{i,1} \xleftarrow{\$} \mathbb{G}$.
$\text{Commit}(m, (s, \mathfrak{t})) \to (x, \mathsf{c}, (\pi_{A_1}, \pi_{A_2}, \pi_B))$: $y = g^x$.
$(A_{(j,1)})_j = \text{Com}(x, s, (g_{j,0}, g_{j,1})_j)$, $(A_{(i,2)})_i = \text{Com}(x, m, (f_{i,0}, f_{i,1}))_i$,
$(B_{(i,0)})_i = \text{Com}(x, \mathfrak{t}, (h_{i,0}, h_{i,1})_i)$, $(B_{(i,1)})_i = \text{Com}(x, \mathfrak{t}, (h_{i,0}, h_{i,1})_i)$.
π_{A_1} proves the knowledge of x s.t. for each j, $y = g^x$ AND $\{A_{(j,1)} = g^x_{j,0}$ OR $A_{(j,1)} = g^x_{j,1}\}$.
π_{A_2} proves the knowledge of x s.t. for each i, $y = g^x$ AND $\{A_{(i,2)} = f^x_{i,0}$ OR $A_{(i,2)} = f^x_{i,1}\}$.
π_B proves the knowledge of x s.t. for each i,
$y = g^x$ AND $\{\{B_{(i,0)} = h^x_{i,0}$ AND $B_{(i,1)} = h^x_{i,1}\}$ OR $\{B_{(i,0)} = h^x_{i,1}$ AND $B_{(i,1)} = h^x_{i,0}\}\}$.
$\text{VerCommit}(\mathsf{c}, (\pi_{A_1}, \pi_{A_2}, \pi_B)) \to 0/1$: Verify the zero-knowledge proofs π_{A_1}, π_{A_2} and π_B.
$\text{Open}(x, \mathsf{c}) \to (\mathsf{m}, \pi)$: $\mathsf{m} = \text{Ope}(x, (A_{(i,2)})_i, (f_{i,0}, f_{i,1})_i)$.
π proves the knowledge of x s.t. $y = g^x$ AND $\Pi^{\ell_2}_{i=1} A_{(i,2)} = \left(\Pi^{\ell_2}_{i=1} f_{i,\mathsf{m}[i]}\right)^x$.
$\text{VerOpen}(\mathsf{c}, \mathsf{m}, \pi) \to 0/1$: Verify the zero-knowledge proof π.
$\text{OpenLDP}(x, \mathsf{c}, (\hat{\mathfrak{s}}, \hat{\mathfrak{t}})) \to (\hat{\mathsf{m}}, \hat{\pi})$: $\mathsf{s} = \text{Ope}(x, (A_{(j,1)})_j, (g_{j,0}, g_{j,1})_j)$,
$\mathsf{m} = \text{Ope}(x, (A_{(i,2)})_i, (f_{i,0}, f_{i,1}))_i$, $\mathfrak{t} = \text{Ope}(x, (B_{(i,0)})_i, (h_{i,0}, h_{i,1})_i)$, $\hat{\mathsf{m}} = \begin{cases} \mathsf{m} & \text{if } \mathsf{s} \oplus \hat{\mathsf{s}} = 0 \\ \mathfrak{t} \oplus \hat{\mathfrak{t}} & \text{else} \end{cases}$
$\hat{\pi}$ proves the knowledge of x s.t.
$y = g^x$ AND $\left(\{\Pi^{\ell_1}_{j=1} A_{(j,1)} = \left(\Pi^{\ell_1}_{j=1} g_{j,\hat{\mathfrak{s}}[j]}\right)^x$ AND $\Pi^{\ell_2}_{i=1} A_{(i,2)} = \left(\Pi^{\ell_2}_{i=1} f_{i,\hat{\mathsf{m}}[i]}\right)^x\}$
OR $\{\Pi^{\ell_1}_{j=1} A_{(j,1)} \neq \left(\Pi^{\ell_1}_{j=1} g_{j,\hat{\mathfrak{s}}[j]}\right)^x$ AND $\Pi^{\ell_2}_{i=1} B_{(i,\hat{\mathsf{m}}[i])} = \left(\Pi^{\ell_2}_{i=1} h_{i,\hat{\mathfrak{t}}[i]}\right)^x\}\}$
$\text{VerOpenLDP}(\mathsf{c}, \hat{\mathsf{m}}, \hat{\pi}, (\hat{\mathfrak{s}}, \hat{\mathfrak{t}})) \to 0/1$: Verify the zero-knowledge proof $\hat{\pi}$.

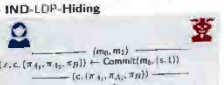
Remark that $h^x_{i,\hat{\mathfrak{t}}[i]} = \begin{cases} h^x_{i,0} & \text{if } \mathfrak{t}[i] = \hat{\mathfrak{t}}[i] \\ h^x_{i,1} & \text{if } \mathfrak{t}[i] \neq \hat{\mathfrak{t}}[i] \end{cases} = \begin{cases} B_{(i,0)} & \text{if } \mathfrak{t}[i] = \hat{\mathfrak{t}}[i] \\ B_{(i,1)} & \text{if } \mathfrak{t}[i] \neq \hat{\mathfrak{t}}[i] \end{cases} = B_{(i,\mathfrak{t}[i] \oplus \hat{\mathfrak{t}}[i])} = B_{(i,\hat{\mathsf{m}}[i])}$.

Security Properties

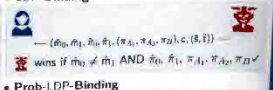
- Classical **Hiding** and **Binding**
- **ROS**-LDP-**Hiding**



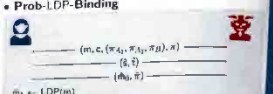
- **IND**-LDP-**Hiding**



- **LDP-Binding**



- **Prob**-LDP-**Binding**



$\text{OptiGuess}(\lambda, \text{LDP}, \hat{m}, m_0, m_1)$:
$b \xleftarrow{\$} \{0, 1\}$
if $\Pr[\hat{m} \leftarrow \text{LDP}(m_0)] > \Pr[\hat{m} \leftarrow \text{LDP}(m_1)]$ then return 0
if $\Pr[\hat{m} \leftarrow \text{LDP}(m_0)] < \Pr[\hat{m} \leftarrow \text{LDP}(m_1)]$ then return 1
if $\Pr[\hat{m} \leftarrow \text{LDP}(m_0)] = \Pr[\hat{m} \leftarrow \text{LDP}(m_1)]$ then return b

Acknowledgments

This work was supported by the France 2030 Projects PEPR Tracia (22-PESN-0006), PEPR iPoP (22-PECY-0002) and AMI CMA CyberINSA (23-CMAS-0019).


CANoé : L'application au service de votre santé vasculaire
Florian Pécune, Solène Jouan, Etienne de Sevin, Camille Maiano, Julien Coelho, Romain Bourcier, Pierre Philip, Richard Redon
Université de Bordeaux, CNRS, SANPSY, UMR 6033, F-33076 Bordeaux, France
Nantes Université, CHU Nantes*, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
Sleep Medicine Service, University Hospital Pellegrin, F-33076 Bordeaux, France
Florian PECUNE
NEUROVASC
Contexte
• La majorité des anévrismes intracrâniens (AIC) sont détectés au cours d'un examen de routine
• Les patients atteints d'un AIC présentent une augmentation de l'anxiété par rapport à la population générale [1,2,3].
Objectif: Accompagner les patients diagnostiqués d'un anévrisme intracrânien non-rompu
• Une interface dédiée à informer et à soutenir les patients, visant à renforcer leur niveau d'éducation thérapeutique afin de mieux comprendre les anévrismes intracrâniens (AIC) et la notion de risque de rupture.
• Un compagnon numérique accompagnant le patient suite à l'annonce médicale et permettant le suivi des facteurs de risque de rupture : consommations d'alcool/tabac/cannabis, mais également sommeil et santé mentale du patient
Suivi sommeil / santé mentale
Phase 1
Questionnaires de suivi
A
Phase 2
Agenda sommeil / sport (7 jours)
A
+
Questionnaires de suivi
B
Phase 2
Agenda sommeil / sport (10 jours)
A
+
Questionnaires de suivi
B
+
Conseils
C
CANoé
Suivi tabac / alcool / cannabis
Agenda Objectifs / Consommations
Bilan suivi consommations
FAQ Anévrisme
Qu'est ce qu'un anévrisme ?
Quels sont les dangers d'un anévrisme?
Comment vivre avec un anévrisme?
Inserm
cnrs

Salman ALMUHAMMAD ALALI

Near-Term Prediction of Ventricular Arrhythmias from Implantable Cardioverter Defibrillator Time-Series Data – A Proof-of-Concept Study

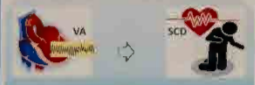
S.A. Alali [1], A. Kachenoura [1], P. Ghoufi [2], G. Carrault [1], S. Boveda [3], R. Garcia [4], W. Aoudjeghout [2,5], F. Kerkouri [2,3,6], A.I. Hernandez [1], E. Marijon [2,5,7], A. Karfoul [1]

[1] Univ Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France | [2] APHP/Inserm, France | [3] Clinique Pasteur | [4] Service de Cardiologie; CHU de Poitiers | [5] INSERM U970 – Hôpital Européen Georges Pompidou | [6] University hospital of Brest; Univ Brest, Laboratoire ORPHY EA 4324, F-29200 Brest, France | [7] European Georges Pompidou Hospital

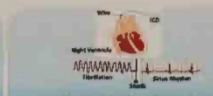
salman.almuhammad-alali@univ-rennes.fr

Context & Objective

Clinical Context

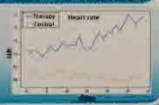


Ventricular Arrhythmia (VA)
a major health concern cause of Sudden Cardiac Death (SCD)

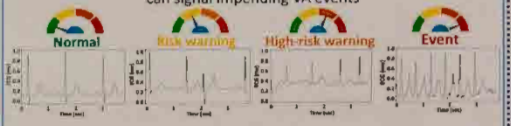


Implantable Cardioverter Defibrillators (ICDs)
Increase fibrillation in high-risk patients

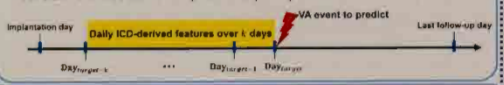
ICDs daily monitor physiological data
offering early warning of arrhythmic events



Hypothesis: Short-term physiological changes in daily telemetry can signal impending VA events

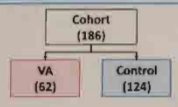


ICD-derived features over past k days → prediction of VA occurrence



Objective: developing and improving short-term VA prediction (**early-detection**) models using daily collected ICD-derived data

Dataset



❑ **Daily recorded variables**

1. Mean ventricular rate
2. Mean ventricular rate at rest
3. Right-Ventricular (RV) lead impedance
4. Shock impedance
5. Mean sensing threshold
6. Minimum sensing threshold
7. RV pacing percentage
8. Physical activity level
9. Supraventricular tachycardia count
10. Non-sustained VT* count
11. VT* in Zone1 count

*VT = Ventricular Tachycardia

Methodology

❑ **Mixed Linear Model (MLM) [1]:** extract temporal trends from repeated measurements

- MLM: $x_i(t) = b_{0i} + b_{1i} \times t + \varepsilon_i(t)$
  - $x_i(t)$: observation for patient i on day t,
  - $b_{0i} = \beta_0 + u_{0i}$: intercept (population-level fixed + patient-specific deviation)
  - $b_{1i} = \beta_1 + u_{1i}$: slope (population-level fixed + patient-specific deviation)
  - $\varepsilon_i(t)$: residual error at day t for patient i

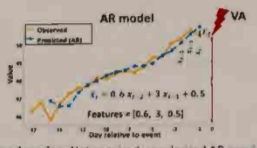
- **4 features per variable:** (11 variables → 44 features)
  - **Raw value** ($Day_{target-1}$)
  - **Slope** (from mixed model)
  - **Difference** ($Day_{target-1}$ - Day_0)
  - **Intercept** (from mixed model)

- **Features engineering:** Morris sensitivity analysis
  44 features → **top 10 features**

❑ **AutoRegressive (AR) model:** captures dynamics over time and encodes them into learnable parameters

- AR(p): $x_i(t) = \sum_{k=1}^{p} \phi_j x_i(t-j) + \varepsilon_i(t)$
  - $x_i(t)$: observed value for patient i on day t
  - $x_i(t-j)$: observed value j days before t (lag j) for patient i
  - ϕ_j: autoregressive coefficient for lag j (persistence/weight)
  - $\varepsilon_i(t) \sim N(0, \sigma^2)$: random noise at time t, with σ^2 noise variance

- **3 features per variable:** (11 variables → 33 features) AR(2) model parameters (ϕ_1, ϕ_2, σ^2) as features



- **Features engineering:** Not required, regularized AR provides inherent selection, retaining the most relevant information

Classifier as predictor:
XGBoost Gradient Boosting with 10-fold cross-validation

Results

❑ **Evaluation criteria**

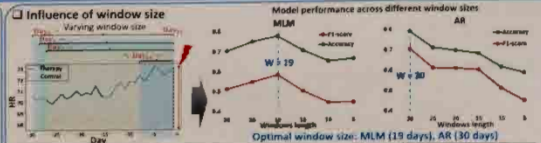
$Se = \frac{TP}{TP+FN}$, $Sp = \frac{TN}{TN+FP}$,

$F1-Score = 2 \times \frac{P^+ \times Se}{P^+ + Se}$,

$Acc = \frac{TP+TN}{TP+TN+FP+FN}$; here, $P+ = \frac{TP}{TP+FP}$

TP: True Positives, **TN**: True Negatives, **FP**: False Positives, **FN**: False Negatives, **Se**: Sensitivity **Sp**: Specificity **P+**: Precision **Acc**: Accuracy

❑ **Influence of window size**



Model performance at the optimal window size for predicting VA on $Day_{target-1}$

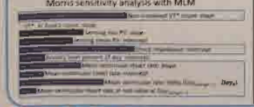
	Acc	Sen	Sp	P+	F1-Score
MLM	0.75	0.532	0.853	0.634	0.578
AR	0.788	0.677	0.854	0.732	0.703

AR-based features improved Sensitivity of VA-risk prediction

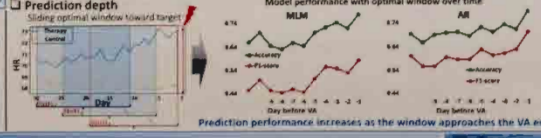
Better sensitive → More lives saved

Better overall performance with AR: F1 0.703, compared to 0.578 for MLM

❑ **Feature importance:** top significant features
Morris sensitivity analysis with MLM



❑ **Prediction depth**



Prediction performance increases as the window approaches the VA event

Conclusion

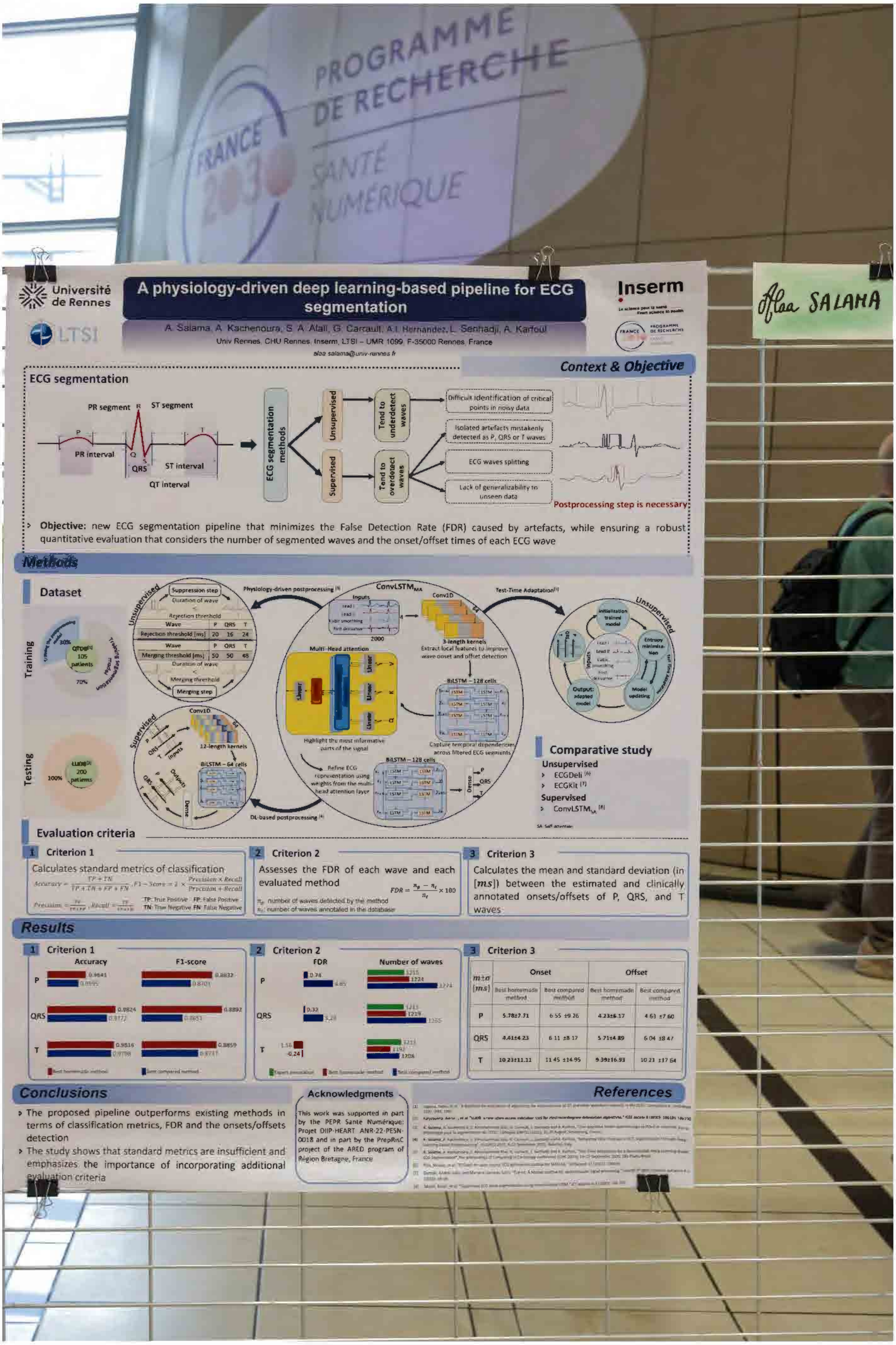
- Proof-of-concept study → feasibility of predicting VAs occurrence using ICD-derived data [1]
- Ventricular rate and the number of non-sustained ventricular tachycardia episodes associated as risk factors [1]
- Shock impedance associated as a potential risk factor → decreased in the documented VA groups [2]
- AR-based approach yields a 21.6% relative F1-score gain in VA prediction over MLM

References

[1]. Ghoufi P., Kachenoura A., Carrault G., Boveda S., Garcia R., Aoudjeghout W., Karfoul F., Marijon E., Karfoul A. "Near-Term Prediction of Ventricular Arrhythmias from Implantable Cardioverter Defibrillator Time-Series Data – A Proof-of-Concept Study". In: Computing in Cardiology (CinC) '52 (2025)

[2]. Sandor C. "Predicting malignant ventricular arrhythmias using real-time remote monitoring." Journal of the American College of Cardiology 81 (2023)

Acknowledgment: this study was funded by the PEPR Santé Numérique, Projet DIIP-HEART



PROGRAMME DE RECHERCHE
FRANCE 2030
SANTÉ NUMÉRIQUE
Université de Rennes
LTSI
A physiology-driven deep learning-based pipeline for ECG segmentation
A. Salama, A. Kachenoura, S. A. Atali, G. Carrault, A. I. Hernandez, L. Senhadji, A. Karfoul
Univ Rennes, CHU Rennes, Inserm, LTSI – UMR 1099, F-35000 Rennes, France
Inserm
Alaa SALAMA
Context & Objective
ECG segmentation
PR segment
ST segment
PR interval
QRS
ST interval
QT interval
ECG segmentation methods
Unsupervised
Supervised
Tend to underdetect waves
Tend to overdetect waves
Difficult identification of critical points in noisy data
Isolated artefacts mistakenly detected as P, QRS or T waves
ECG waves splitting
Lack of generalizability to unseen data
Postprocessing step is necessary
Objective: new ECG segmentation pipeline that minimizes the False Detection Rate (FDR) caused by artefacts, while ensuring a robust quantitative evaluation that considers the number of segmented waves and the onset/offset times of each ECG wave
Methods
Dataset
Training
Testing
Physiology-driven postprocessing
ConvLSTM
Test-Time Adaptation
Suppression step
Merging step
Multi-Head attention
BiLSTM – 128 cells
DL-based postprocessing
Comparative study
Unsupervised
ECGDeli
ECGKit
Supervised
ConvLSTM
Evaluation criteria
Criterion 1
Calculates standard metrics of classification
Criterion 2
Assesses the FDR of each wave and each evaluated method
Criterion 3
Calculates the mean and standard deviation (in [ms]) between the estimated and clinically annotated onsets/offsets of P, QRS, and T waves
Results
Criterion 1
Accuracy
F1-score
Criterion 2
FDR
Number of waves
Criterion 3
Onset
Offset
Conclusions
The proposed pipeline outperforms existing methods in terms of classification metrics, FDR and the onsets/offsets detection
The study shows that standard metrics are insufficient and emphasizes the importance of incorporating additional evaluation criteria
Acknowledgments
This work was supported in part by the PEPR Santé Numérique: Projet DIIP-HEART ANR-22-PESN-0018 and in part by the PrepRisC project of the ARED program of Région Bretagne, France
References

Sofia KAISARIDI

PROGRAMME DE RECHERCHE
SANTÉ NUMÉRIQUE

LEArning Spatiotemporal patterns in PYthon

Sofia Kaisaridi[1], Juliette Ortholand[1], Caglayan Tuna[1,2], Gabrielle Casimiro[1], Stanley Durrleman[1], Sophie Tezenas Du Montcel[1]
1 Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
2 Inria, Université Paris Cité, Inserm, HeKA, F-75015 Paris, France

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0016

ARAMIS LAB
BRAIN DATA SCIENCE

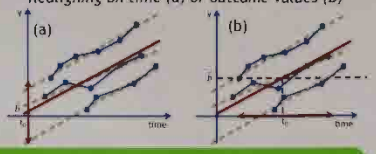
Introduction: challenges from chronic diseases

Disease progression models are prominent tools for unravelling the mechanisms of chronic diseases. Yet they face different challenges:

- **Heterogeneity:** the model must account for inter-individual variability
- **Partial trajectories:** patients follow-up only covers a short part of the disease
- **Realigning observations:** patients have different speeds of progression and age at onset

To achieve these goals, we introduce Leaspy, an open-source library in Python, which implements spatio-temporal models for disease progression.

Realigning on time (a) or outcome values (b)

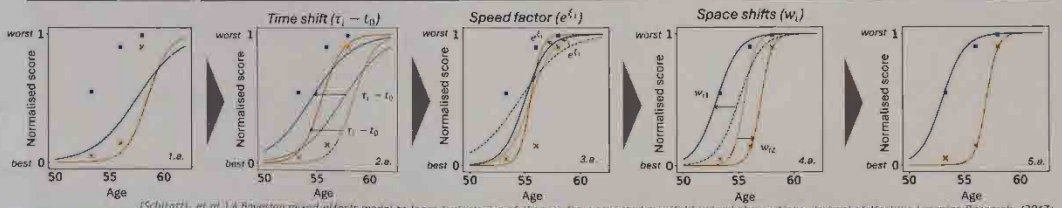


Methods: the model behind Leaspy

Leaspy is built on the Spatiotemporal model proposed by [Schiratti et al.]: a non-linear mixed-effects model. It realigns patients on a common disease timeline (temporal aspect of the disease) thanks to a latent disease age. The remaining heterogeneity between individuals, which is reflected in a variability in the reordering of outcome progression (spatial aspect) is then captured.

The models' parameters account for these temporal and spatial effects, and their estimation is performed with an MCMC-SAEM algorithm.

Population progression	**Temporal** Random effects	**Spatial** Random effects	**Individual** progression



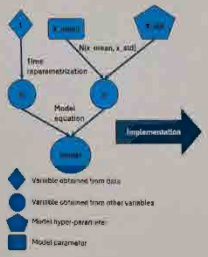
[Schiratti, et al.] A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. Journal of Machine Learning Research (2017)

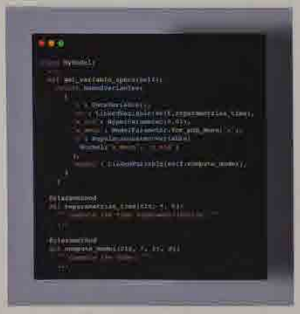
A user-friendly interface

With a simple API, users can easily tailor their model for immediate use. This open-source package is developed following modern practices (systematic testing, continuous integration and deployment, peer-review procedures).

Implement your own model

Leaspy is a flexible library that allows the user to implement their own model.



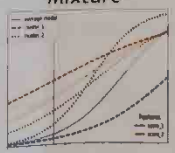


Applications

Leaspy has been successfully applied to various diseases (Alzheimer's, Parkinson's, Huntington's, ALS, Ataxia, ...) and allows fitting multivariate models to different types of input (continuous, ordinal or time-to-event in a joint model), potentially adjusting trajectories for covariates and then making dynamic predictions for new patients. In June 2020, it ranked 2nd for cognitive decline prediction in the TADPOLE challenge *[Nguyen, et al]*.

Different models ...

Mixture

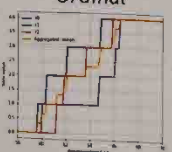


(Kaisaridi et al., ISCB 2025)

Covariates

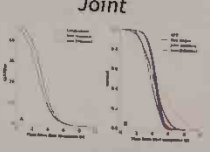
(Fournier, MICCAI 2023 Casimiro, ISCB 2025)

Ordinal



(Poulet et al., Stat in Med, 2023)

Joint



(Ortholand et al., arxiv, 2024)

... for different applications

Patient's selection

(Maheux et al., Nature Com. 2023)

Multi-dimension

(Kaisaridi et al., Neurology, 2025)

[Nguyen, M. et al.] Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage (2020).

SORBONNE UNIVERSITÉ · Paris Brain Institute ICM · Inserm · CNRS · Inria · ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

A REGULARIZED MULTI-STATE MODEL FOR COVARIATE SELECTION WITH INTERVAL-CENSORED SURVIVAL DATA

Ariane Bercu[1], Agathe Guilloux[2], Cécile Proust-Lima[1], Hélène Jacqmin-Gadda[1]

[1]Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France
[2]Inria Paris F-75015, INSERM, CRC, Univ. Sorbonne Paris, Sorbonne Université, F-75006 Paris, France
ariane.bercu@u-bordeaux.fr, agathe.guilloux@inria.fr, cecile.proust-lima@u-bordeaux.fr, helene.jaqmin-gadda@u-bordeaux.fr

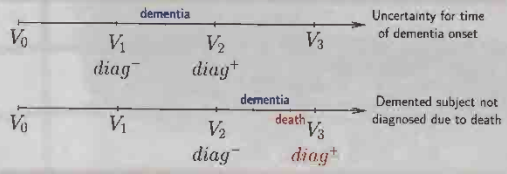
MOTIVATING APPLICATION

Predict the **risk of dementia in a cohort of elderly subjects (3C Study)** according to a large number of possible predictors (age, sex, history of disease and treatment, cognitive testing, dependency scaling, brain volumes, blood biomarkers, genetics...).

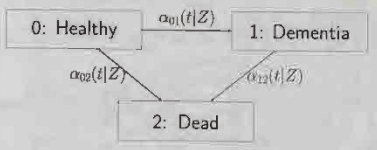
Challenges:

- interval censoring of dementia onset (diagnosis visit every 3y in 3C)
- competing risk of death (death date collected continuously in 3C)

INTERVAL CENSORING WITH COMPETING RISK OF DEATH



ILLNESS-DEATH MODEL (IDM)



Aim: Adapt variable selection methods using Elastic Net regularization to the illness death model for interval censored time of illness onset (Reg IDM-ICT).

METHODS

θ : baseline intensity parameters β : regression parameters

$$\alpha_{01}(t|Z)=\alpha_{0,01}(t,\theta_{01})e^{\beta_{01}'Z} \quad \alpha_{02}(t|Z)=\alpha_{0,02}(t,\theta_{02})e^{\beta_{02}'Z} \quad \alpha_{12}(t|Z)=\alpha_{0,12}(t,\theta_{12})e^{\beta_{12}'Z}$$

$$\ell_{\text{pen}}(\theta,\beta,a,\lambda) = \ell(\theta,\beta) - \text{pen}(\beta_{01},a,\lambda_{01}) - \text{pen}(\beta_{02},a,\lambda_{02}) - \text{pen}(\beta_{12},a,\lambda_{12})$$

where $\text{pen}(\beta_{hl},a,\lambda_{hl})$ the elastic net penalty on $h \to l$ given by $\lambda_{hl}a\|\beta_{hl}\|_1 + \lambda_{hl}(1-a)\|\beta_{hl}\|_2^2$, with $\lambda_{hl} > 0$ and $0 \le a \le 1$.

ESTIMATION PROCEDURE OF REG IDM-ICT

Outer loop

From a predefined grid of hyper-parameters $(a,\lambda_{01},\lambda_{02},\lambda_{12})$

Inner loop

Maximization of the regularized log-likelihood for fixed penalty parameter

$$(\widehat{\theta}(a,\lambda),\widehat{\beta}(a,\lambda)) = \arg\max_{\theta,\beta}\{\ell(\theta,\beta) - \text{pen}(\beta,a,\lambda)\}$$

using a 2-step algorithm combining cyclical gradient descent and Newton like algorithm.

Calculate the bayesian information criterion (BIC),

$$BIC(a,\lambda) = -2\ell(\widehat{\theta}(a,\lambda),\widehat{\beta}(a,\lambda)) + \log(\#subjects)\|\widehat{\beta}\|_0.$$

Overall, minimize the BIC $\quad (\widehat{\theta}(\hat{a},\hat{\lambda}),\widehat{\beta}(\hat{a},\hat{\lambda})) = \arg\min_{a,\lambda}\{BIC(a,\lambda)\}$

SIMULATION STUDY

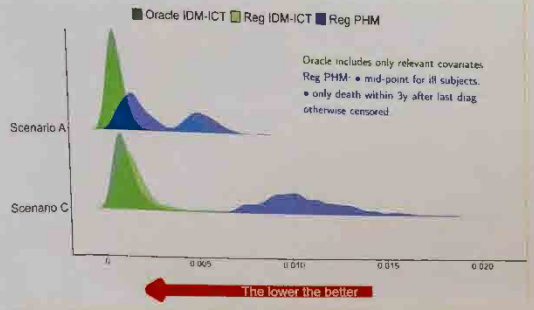
Low events rate and frequent visit A — High event rates and sparse visit

50 / 200 among 300 / 600 demented subjects died before diagnosis.

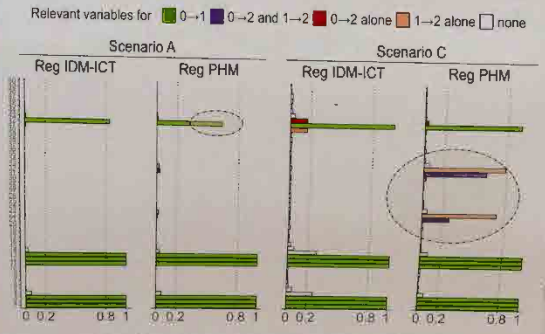
50 correlated centered Gaussian covariates on each transition

PREDICTIVE ABILITY

Mean Square Error on the Prediction (MSEP) for a healthy subject of having dementia up to t, the time of first event between the time of dementia onset, death and censoring.



VARIABLE SELECTION ON THE TRANSITION TO ILLNESS



Variable selection frequency on transition 0→1

CONCLUSION

Neglecting IC may lead to incorrect selection as dementia predictors of variables only associated with death and large error of prediction.

The proposed method implemented in the package github.com/arianebercu/HIDeM corrects these issues with performance close to the oracle model.

REFERENCES

P. Joly et al. (2002). Biostatistics, 3, 433–443
K. Leffondré et al. (2013). International Journal of Epidemiology, 42, 1177–1186
N. Simon et al. (2011). Journal of Statistical Software, 39
H. Zou et al. (2005). JRSSS B. Statistical Methodology, 67, 301–320.
V. Philipps et al. (2021). The R Journal, 13, 273

Qualification & gestion des données manquantes dans les modèles d'aide à la décision clinique en vie réelle : revue préliminaire de la littérature avec PRISMA

Pauline FRACASSO, Sandie CABON, Morgane PIERRE-JEAN, Gouenou COATRIEUX, Marc CUGGIA

Univ Rennes, CHU Rennes, INSERM, LTSI - UMR 1099, F-35000 Rennes, France

INTRODUCTION

Les entrepôts de données de santé permettent la réutilisation des données produites dans le cadre du soin et la mise en place d'outils d'aide à la décision clinique [1]. Cependant, la qualité des données est souvent altérée par **des données manquantes (DM)**, issues de processus de collecte non systématiques [2]. Ces absences peuvent être définies selon i) différents **mécanismes** : **MCAR** (DM totalement aléatoires), **MAR** (DM dépendantes de variables observées) ou **MNAR** (DM dépendantes de variables non-observées, en raison du processus de collecte lui-même) [2], ii) La **proportion** de données manquantes [2] et iii) le **motif** que les données manquantes suivent [3].

Dans un contexte où l'IA Act met l'accent sur la fiabilité des modèles à base d'apprentissage, l'absence d'un cadre méthodologique standardisé pour qualifier et correctement gérer les données manquantes constitue un frein majeur. Une mauvaise gestion des données manquantes peut introduire des biais significatifs, compromettant ainsi leur fiabilité. Cette étude vise à **recenser les méthodes existantes** pour **qualifier** et **gérer** ces DM, dans le but de **concevoir une chaîne de traitement automatique**, optimisant ainsi la qualité des données cliniques et renforçant la confiance dans les modèles d'IA en santé.

MATÉRIEL & MÉTHODE

Dans cette revue, nous abordons la question suivante : **"Existe-t-il un cadre permettant de qualifier les données manquantes et de sélectionner la méthode de gestion optimale ?"**

Nous y répondons avec 3 questions de recherche :

1. Quelles sont les méthodes de qualification des données manquantes et quelles sont les conditions d'application ?
2. Quelles sont les méthodes de gestion des données manquantes et quelles sont les conditions d'application ?
3. Quels articles établissent un lien entre la qualification des données manquantes et les méthodes de gestion des données manquantes ?

LIEN ENTRE QUALIFICATION ET GESTION DES DONNÉES MANQUANTES

Etape PRISMA :

1 EQUATION DE RECHERCHE

("missing data" OR "data missingness" OR "incomplete data" OR "missing entries")
AND
(("characterization" OR "assessment" OR "evaluation" OR "analysis" OR "detection" OR "profiling" OR "monitoring" OR "qualification")
OR
("imputation" OR "management" OR "handling" OR "remediation" OR "intervention" OR "treatment")

2 IDENTIFICATION

Records identified through database searching (n=64)

PUBMED (34)
WEB OF SCIENCE (30)

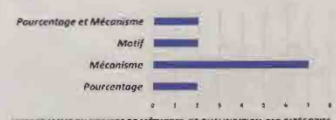
3 SELECTION

Records after duplicates Removed (n=52)

Records screened (n=24)
Removed :
(Unavailable (n=8),
Not a review (n=5),
Lack of comparison of the methods' characteristics (n=15))

RESULTATS

1 MÉTHODES DE QUALIFICATION



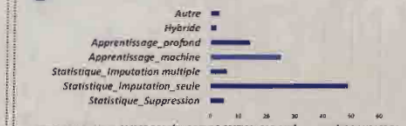
HISTOGRAMME DU NOMBRE DE MÉTHODES DE QUALIFICATION PAR CATÉGORIES CITÉES DANS LES 24 ARTICLES

Méthode	Catégorie	% d'article et nombre (n) qui cite la méthode	Exemple
MISSING DATA PATTERN	Motif	17% (n=4)	[4]
SENSITIVE ANALYSIS	Mécanisme	17% (n=4)	[5]
LITTLE'S TEST	Mécanisme	13% (n=3)	[6]

TABLEAU DES MÉTHODES CITÉES DANS AU MOINS 3 ARTICLES

13 MÉTHODES DE QUALIFICATION DE DM RENCENSÉES AU TOTAL

2 MÉTHODES DE GESTION



HISTOGRAMME DU NOMBRE DE MÉTHODES DE GESTION PAR CATÉGORIE CITÉES DANS LES 24 ARTICLES

Méthode	Catégorie	% d'article et nombre (n) qui cite la méthode	Exemple
MICE	Statistique imputation seule	88% (n=23)	[7]
CCA	Statistique suppression	71% (n=17)	[8]
MEAN IMPUTATION	Statistique imputation seule	46% (n=11)	[9]
LOCF	Statistique imputation seule	46% (n=11)	[10]
EM	Statistique imputation seule	42% (n=10)	[11]
IPW	Statistique imputation seule	29% (n=7)	[12]
ACA	Statistique suppression	21% (n=5)	[13]
KNN	Apprentissage machine	21% (n=5)	[14]

TABLEAU DES MÉTHODES CITÉES DANS AU MOINS 5 ARTICLES

104 MÉTHODES DE GESTION DE DM RENCENSÉES AU TOTAL

3 QUALIFICATION ET GESTION

SUR LES 24 ARTICLES

- Aucun article ne recense uniquement des méthodes de qualification de DM
- 9 articles recensent uniquement des méthodes de gestion de DM

→ **13 articles relient méthodes de qualification et méthodes de gestion de DM**

SUR LES 13 ARTICLES

METHODES DE QUALIFICATION DES DM COUVERTES : **100% (n=13)**

METHODES DE GESTION DES DM COUVERTES : **58% (n=60)**

CONCLUSION & PERSPECTIVES

- Cette revue a permis un premier recensement des approches existantes pour la qualification et la gestion des données manquantes, soulignant leur dépendance à des critères de qualification encore peu standardisés. Avec 104 méthodes de gestion recensées contre seulement 13 de qualification, une seconde revue est en cours pour élargir le spectre des méthodes de qualification.
- À terme, ces travaux viseront à concevoir une chaîne de traitement automatique, optimisant la gestion des données manquantes en santé selon leur qualification, en amont des applications d'aide à la décision clinique.

RÉFERENCES

[1] Meystre S.M et al. Clinical Data Reuse or Secondary Use: Current Status and Potential Future. Progress. Yearbook of Medical Informatics. 2017

[2] Austin P.C et al. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. The Canadian Journal of Cardiology. 2021

[3] Little R.J. et al. Missing Data Analysis. ANNUAL REVIEW OF CLINICAL PSYCHOLOGY. 2024

Accédez aux références complètes via le QR code

CHU rennes · UNIVERSITÉ DE RENNES 1 · LTSI · TRACIA · Inserm

Assessing bioinformatics software annotations: bio.tools case study

Ulysse LE CLANCHE[1], Sarah COHEN BOULAKIA[2], Yann LE CUNFF[1], Olivier DAMERON[3] and Alban GAIGNARD[3,4]
[1] Université Rennes, Inria, CNRS, IRISA - UMR 6074, Rennes 35000, France
[2] Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91405, Orsay, France
[3] Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France
[4] IFB-core, Institut Français de Bioinformatique (IFB), CNRS, INSERM, INRAE, CEA, 91057 Evry, France
Corresponding author:
Projet financé dans le cadre de France 2023 par l'ANR sous la référence ANR-22-PESN-0007 ShareFAIR

Ulysse LE CLANCHE

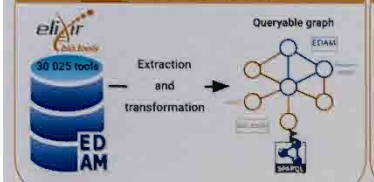
Introduction

- Specialized registries in the life sciences, such as TeSS, bio.tools, and WorkflowHub, support the findability and interoperability of training materials, software tools, and analysis pipelines
- The EDAM ontology :
  - provides shared vocabulary for data types, formats, operations, and topics,
  - enables consistent annotation and facilitates tasks such as workflow composition and data retrieval
- Bio.tools hosts over 30,000 annotated software entries.

Challenges: assess the reliability and precision of software annotations.

Material and methods

Bio.tools dataset



The EDAM ontology

Branche EDAM	Number of annotations in bio.tools
Data	11 K
Format	10 K
Operation	70 K
Topic	100 K

EDAM

The study focused on the **operations** and **topics** branches, covering most tools.

Shannon entropy

p represents the probability of occurrence of an annotation a

$$H(a) = -p(a) \times \log_2(p(a))$$

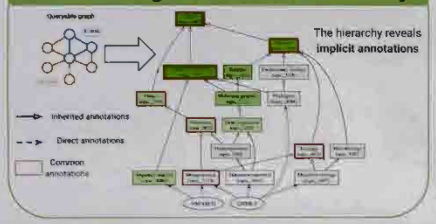
Near 0 for rare annotations — Near 0 for frequent annotations

$$H(\text{tool}) = \sum_{a \in \text{annot}(\text{tool})} H(a)$$

Shannon entropy (H) was used to quantify **the degree of information in annotations.**
Entropy is low for rare or frequent annotations.
Entropy is higher for intermediate annotations.

Results

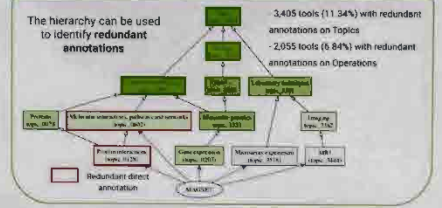
1. Considering the annotation hierarchy



The hierarchy reveals **implicit annotations**

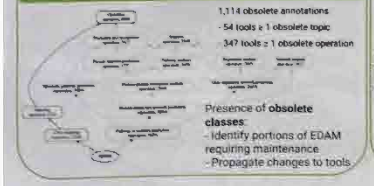
2. Not all direct annotations are needed

The hierarchy can be used to identify **redundant annotations**



- 3,405 tools (11.34%) with redundant annotations on Topics
- 2,055 tools (6.84%) with redundant annotations on Operations

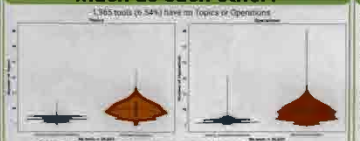
3. All direct annotations are not relevant



- 1,114 obsolete annotations
- 54 tools ≥ 1 obsolete topic
- 347 tools ≥ 1 obsolete operation

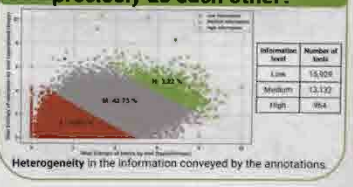
Presence of **obsolete classes**:
- Identify portions of EDAM requiring maintenance
- Propagate changes to tools

4. Are all the tools annotated as much as each other?



After filtering, the hierarchy is taken into account :
- **Heterogeneity in the number of direct/indirect annotations**
- increases the number of annotations

5. Are the tools annotated as precisely as each other?



Information level	Number of tools
Low	15,109
Medium	13,132
High	854

Heterogeneity in the information conveyed by the annotations.

Conclusion

Contributions

- Consider inherited classes in annotation assessment
- Identify redundant and obsolete EDAM annotations
- Identification of poorly annotated and uninformative tools

Future works

- Analysis on Format/Data branch
- Gold standard to better assess bio.tools annotations quality
- Tools to help curation task to improve Bio.tools and the EDAM ontology

Raphaël FOURCHAULT

PSL UNIVERSITÉ PARIS

Valorisation des données du SNDS en sciences sociales

Présentation par Raphaël Fourchault

INSTITUT SANTÉ NUMÉRIQUE EN SOCIÉTÉ

Le contexte

L'accès au Système National des Données de Santé (SNDS) est de plus en plus aisé pour les chercheurs en sciences sociales, notamment via l'accès permanent donné au CNRS en 2021. Le potentiel du SNDS pour les sciences sociales est immense, mais va de pair avec des difficultés d'exploitation. Ces données sont très complexes à manipuler, et font l'impasse sur les caractéristiques socio-démographiques des patients.

La transformation d'une base de gestion en une base de recherche

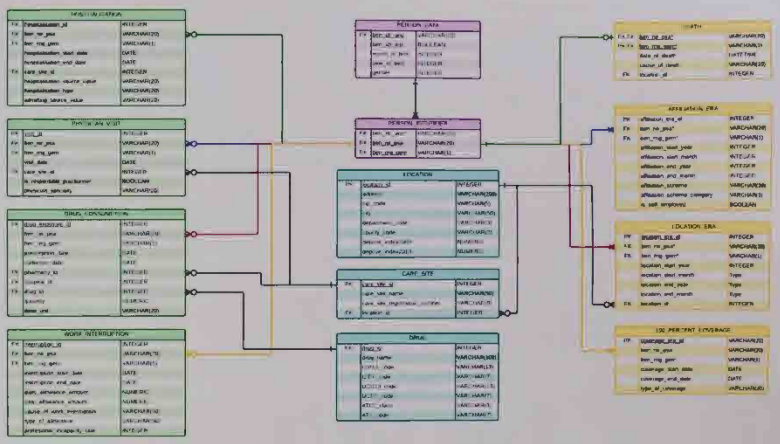
Un travail important est en cours pour établir un **schéma relationnel simplifié**, une documentation commune, et fournir des **codes en open source** permettant de générer des tables prêtes à l'emploi, avec des jointures simplifiées. Ces tables sont construites pour **regrouper l'information de même nature** (séjours hospitaliers, visites médicales en ville, consommation de médicaments, suivi géographique des patients, etc.) qui est actuellement éparse dans le SNDS en raison de **règles de remontée administrative complexes** et de variables non harmonisées selon les régimes d'assurance, les établissements de soin ou les années. Ces tables permettront ainsi aux chercheurs de construire des **échantillons de travail** grâce à des variables moins nombreuses (environ 80 contre plus de 5000 dans le SNDS) et d'étudier les comportements de santé des individus, les pratiques des professionnels de santé, l'organisation des soins sur le territoire, ou encore les liens entre santé et travail.

Nos objectifs sont les suivants :

1. Développer une expertise collective autour du SNDS
2. Apparier pour pallier le manque de données socio-démographiques
3. Mener des travaux en sciences sociales sur le SNDS et ses appariements

1. Le schéma relationnel simplifié et le dépôt Git.

Le schéma simplifié contient des tables de nature médico-administrative : table des visites de professionnels de santé en ville, des hospitalisations, des affections longue durée, des nomenclatures des médicaments en CIP13... La table *hospitalisation* contient par exemple les séjours hospitaliers et renseigne un identifiant patient, les dates et type de séjour, les diagnostics en CIM10, l'établissement visité etc. Les tables sont reliées entre elles par des clés, permettant aussi de joindre des tables extérieures telles que la table des établissements construite à partir des répertoires SIRENE de l'INSEE.



Le dépot Git permettra l'accès aux codes générant les différentes tables requises pour les projets de recherche :

2. L'appariement du SNDS à des données socio-économiques.

- **Un projet d'appariement du SNDS avec SHARE** - la grande enquête européenne sur la santé et le vieillissement menée par Dauphine-PSL - est en cours
- **Les appariements existants sont étudiés et valorisés dans des travaux de recherche** (e.g., EDP-Santé) .

SNDS SHARE insee

3. Quel usage du SNDS et de ses appariements dans le domaine des sciences sociales ?

- **Mesurer l'évolution de la consommation de soins et des dépenses de santé dans le temps**. Identifier le rôle du vieillissement, du progrès technique, de l'offre de soins et de la régulation, ce qui nécessite les données administratives fines du SNDS pour comprendre l'évolution par poste de soin et profil de patient.
- **Comprendre l'impact de l'offre de soins sur la qualité et le recours aux soins.** Lien entre concurrence et qualité du marché hospitalier ; Impact des pénuries (médicaments, professionnels de santé) sur le recours aux soins et évaluation de la perte de bien-être associée.
- **Comprendre les déterminants de la demande de soins.** Quels sont les parcours de soins type des patients entre les différents offreurs de soin (hôpitaux, cliniques, professionnels libéraux, si possible Ehpad) en fonction de leurs caractéristiques sociales ? Quelle adhésion aux politiques vaccinales en fonction des caractéristiques socio-économiques ? Quelle demande de soins suite à des chocs macro-économiques ? Quels appariements médecins-patients et hôpitaux-patients ?

L'équipe

- **Emmanuel Didier** est sociologue, directeur de recherche CNRS au Centre Maurice Halbwachs (ENS-PSL/EHESS). Il est le responsable du projet SaNSo.
- **Mathilde Godard** est économiste, chargée de recherche CNRS au LEDa (Dauphine-PSL). Elle est la responsable scientifique de l'axe "Valeur et Valorisation" du projet SaNSo.
- **Raphaël Fourchault** est ingénieur de recherche au LEDa, Dauphine-PSL.
- **Louis Arnault**, IRDES et SHARE-France.
- **Nicolas Belorgey**, Dauphine-PSL.
- **Cécile Charles**, Ingénieur de recherche, Dauphine-PSL.
- **Brigitte Dormont**, Dauphine-PSL.
- **Clémentine Garrouste**, Dauphine-PSL.
- **Daniel Herrera Araujo**, Dauphine-PSL.
- **Florence Jusot**, Dauphine-PSL.
- **Elsa Perdrix**, Dauphine-PSL.
- **Thomas Renaud**, Ingénieur de recherche, Dauphine-PSL et SHARE-France.

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0004 : Projet SaNSo - Axe "Valorisation et valeur des données de santé"

FRANCE 2030 PROGRAMME DE RECHERCHE SANTÉ NUMÉRIQUE

cnrs

Fabien LAPORTE

Incorporating complex genetic model into risk stratification

F. Laporte[1], H. Perdry[2], A. Herzig[3], C. Dina[1], E. Genin[3], R. Redon[1]

1. Nantes Université, CNRS, INSERM, l'institut du thorax, F-44000 Nantes, France.
2. CESP Inserm, U1018, Université Paris-Saclay, UVSQ, Villejuif, France
3. Inserm, Université de Brest, EFS, CHU Brest, UMR 1078, GGB, F-29200 Brest, France

BACKGROUND

Nowadays polygenic risk scores (PRS) have multiple clinical applications as risk stratification for disease prevention. Most of the PRS algorithm are based on a weighted sum of the effects of risk variants [1]. Genetic architecture of diseases could be more complex therefore classical PRS could perform poorly. Rare variant mutations and epistasis effects have to be accounted for.
Here we present our work to develop the rcRS algorithm, a PRS algorithm accouting for epistasis effects near Genome-wide association study (GWAS) significant hits.

SIMULATING CONTROLS

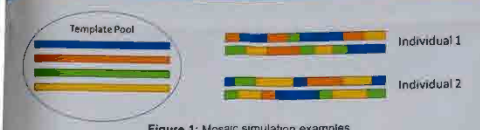


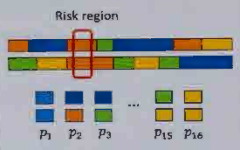
Figure 1: Mosaic simulation examples

Haplotypes are generated in two steps:
1. Mosaic break points are randomly drawn
2. Tiles are randomly drawn in the template pool

Individuals are composed of two haplotypes.
The code is available at https://github.com/genostats/Mozza

SIMULATING CASES

Cases are simulated in three stages:
1. Mosaic break points are randomly drawn
2. Non-at-risk region tiles are randomly drawn
3. Risk region tiles are drawn using probabilities depending of the risk of the possible combinaison of haplotypes



$p_i = r_{h_1 h_2} P(h_1) P(h_2)$ with $r_{h_i h_j}$ risk associated to the combination of haplotypes h_i and h_j and $P(h_i)$ the probability to draw haplotype h_i in the sampling pool [2].

rcRS: Polygenic Risk Score Algorithm accounting for epistasis

To account for possible epistatsis effects within regions of interest, partially linear tree-based regression (GPLTR) [3] are computed within each region of interest. Each GPLTR gives a prediction and those predictions are combined using a generalized linear model.

Prediction thresholds are selected with respect to the AIC of the derived GLM. To avoid computational burden, a grid step iteration is used. At each iteration, only the neighbor combinations (one threshold modified at a time) are computed and the best one is chosen as the starting point of the next iteration. This method converges to the locally best AIC.

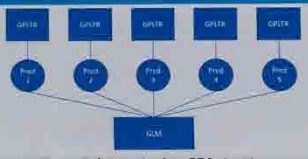


Figure 2: Schematic of the PRS algorithm

RESULTS

2000 cases and 5000 controls were simulated using the previous algorithms for common variants. 4 regions were selected as associated with the disease and the risk associated to the genotype combination « 020 » is 10 (resp. 30) for the region 1, 3 and 4 (resp. 2). Simulation sanity test was performed using GWAS and genotype frequencies (Figure 3).

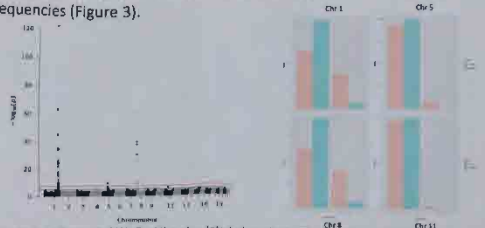


Figure 3: Left panel: GWAS of the simulated phenotype. Right panel: genotype frequencies in each region (top chromosomes 1 and 5, bottom chromosomes 8 and 11)

All regions were either significant or close to significance using the GWAS. Moreover, there is a difference regarding the genotype frequencies between cases and controls.

Once the cohort generated, a 5-fold cross validation was performed to compared the rcRS algorithm to the classical PRS (weighted and GLM). ROC curves are displayed in Figure 4 (left panel).
Then a real dataset with Intracranial Aneurysm carrier is studied (685 cases and 1,064 controls). Again, a 5-fold cross validation was performed to compare rcRS to the GLM PRS. ROC curves are displayed in Figure 4 (right panel).
Regarding the simulated data, the rcRS algorithm perfoms better than its rivals with an increase of AUC. Unfortunately, our first test on real data shows a better predictive ability of the GLM PRS. We will adjust our algorithm parameters and expect an increase of AUC.

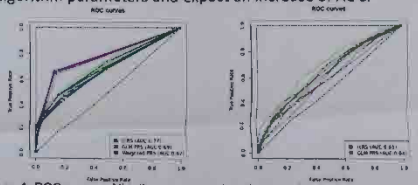


Figure 4: ROC curves of the three compared methods. Left panel: simulated dataset Right panel: real dataset

CONCLUSION

First, the method to simulate a cases/controls cohort using existing genotyping pool with common variations was set up and validated.

Second, accounting for epistasis effects within the PRS algorithm leads to an increase of the AUC regarding simulated data. Further investigation regarding the real dataset are needed.

A method to simulate cases with rare variant mutations [4] is currently tested with convincing results. Then rare variation mutations will be added within the PRS algorithm based on the variant annotations and burden results.

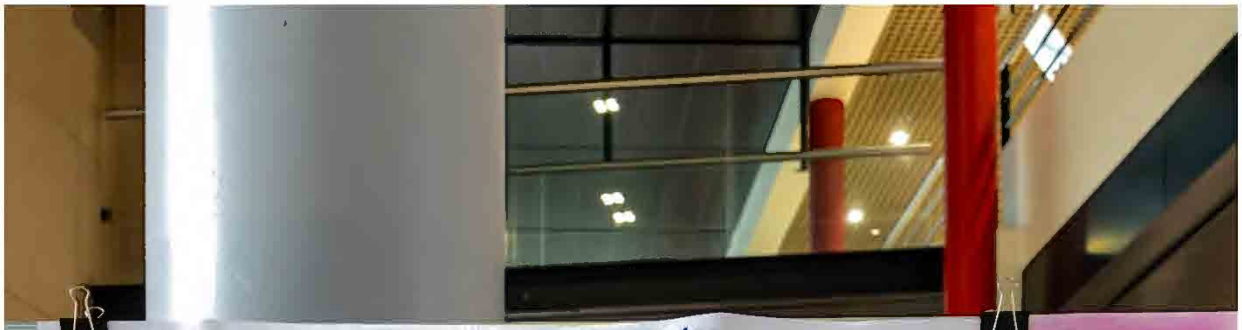
REFERENCES

[1] Euesden, J., Lewis, C. M., & O'reilly, P. F. (2015). PRSice: polygenic risk score software. *Bioinformatics*, *31*(9), 1466-1468
[2] Zhan Su, Jonathan Marchini, Peter Donnelly, HAPGEN2: simulation of multiple disease SNPs, *Bioinformatics*, Volume 27, Issue 16, August 2011, Pages 2304–2305, https://doi.org/10.1093/bioinformatics/btr341
[3] Cyprien Mbogning, Hervé Perdry, Philippe Broët; A Bagged, Partially Linear, Tree-Based Regression Procedure for Prediction and Variable Selection. *Hum Hered* 2 July 2015; 79 (3-4): 182–193. https://doi.org/10.1159/000380850
[4] Clarke, B., Holtkamp, E., Öztürk, H., Mück, M. Wahlberg, M., Meyer, K., ... & Stegle, O. , (2023). Integration of variant annotations using deep set networks boosts rare variant association genetics. *bioRxiv*, 2023-07.

Authors have no conflict of interest

Fabien Laporte, Human Genetics
fabien.laporte@univ-nantes.fr

Project funded by the ANR under the France 2030 program with the reference ANR-22-PESN-0008



Brain Health Trajectories

brain health trajectories

Coordinateur du projet: Viktor Jirsa (INSERM)

WP1 : Jean-François Mangin (CNRS-CEA), Olivier David (INSERM)

WP2 : Demian Battaglia (CNRS), Andrea Brovelli (CNRS)

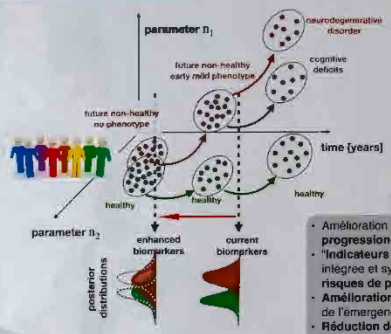
WP3 : Bertrand Thirion (INRIA), Spase Petkoski (INSERM)

Contributeurs : Emmanuel Barbier (INSERM), Mircea Polosan (INSERM-CHUGA)

Spase PETKOSKI

CONTEXTE - POSITIONNEMENT

- Les coûts annuels actuels par patient atteint de la maladie d'Alzheimer (MA) en Europe sont estimés à environ 40 000 euros/an/patient [1].
- Le nombre de patients atteints de la maladie d'Alzheimer devrait passer de 14 millions en 2019 à 50 millions en 2050.
- Les facteurs de risque modifiables sont responsables d'environ 40 % des démences dans le monde [2].
- Le diagnostic au cours de la phase prodromique présymptomatique ou de la déficience cognitive légère (DCL) plutôt qu'au cours de la phase de démence devrait permettre d'économiser environ 8 000 milliards de dollars [3].
- Le diagnostic précoces permet de prédire les trajectoires de la maladie cérébrale et de les modifier.

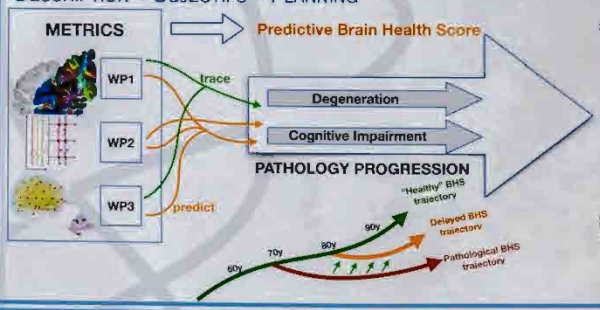


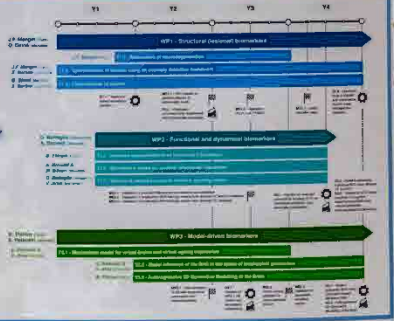
- **Objectif 1** : développer, valider et optimiser une batterie de mesures quantifiant l'état de santé du cerveau d'un individu.
- **Objectif 2** : Démontrer que la quantification de l'état de santé du cerveau est plus performante que les marqueurs individuels et qu'elle peut donc accélérer le diagnostic et la stratification précoce des sujets à risque.

IMPACTS ET RETOMBÉES:

- Amélioration de la capacité de **suivi et anticipation de la progression** d'une pathologie.
- **"Indicateurs augmentés"** simples et compactes qui de façon intégrée et synergique décrivent au niveau personnalisé **risques de progression et chances de résilience cognitive.**
- **Amélioration de la qualité de vie des patients** par la réduction de l'émergence et l'anticipation de troubles cognitifs graves
- **Réduction des coûts** associés au suivi médical et social et à l'assistance aux patients déments.

DESCRIPTION – OBJECTIFS – PLANNING





WP1 - STRUCTURAL BIOMARKERS

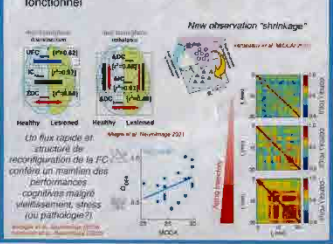
- T1.1 - Scores quantifiant l'ampleur de la neurodégénération fondés sur des abaques normatifs
- T1.2 - Scores quantifiants la charge lésionnel issus d'outils IA de détection d'anomalie (auto-encodeurs)
- T1.3 - Classification automatique des lésions détectées (tumeurs, malformations, post-trauma, post-ischemic, post-inflammation…)

Quantification des atrophies et détection d'anomalies

F-tracts: 1200 patients épileptiques (Europe)

WP2 - DYNAMICAL BIOMARKERS

- T2.1 - Nouveaux méthodes pour l'estimation invariante de la FC, adaptés pour un suivi de données avec variation longitudinale de la structure
- T2.2 - Détection de perturbations précoces de la **communication** par des nouveaux méthodes de **FC dirigée** et de EC appliqués à l'IRMf
- T2.3 - Mesures de la reconfiguration altéré du réseau fonctionnel

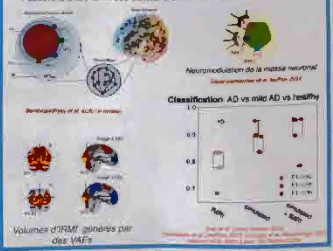


Un flux rapide et structuré de reconfiguration de la FC confère un maintien des performances cognitives malgré vieillissement, stress (ou pathologie?)

WP3 MODEL-DRIVEN BIOMARKERS

- T3.1 Modèle mécaniste pour les cerveaux et les trajectoires virtuelles de vieillissement
- T3.2 Inférence du modèle de la trajectoire BHS dans l'espace des paramètres biophysiques
- T3.3 - Modélisation générative autorégressive 3D de la dynamique cerebrale contrainte par la structure

Passerelle entre le niveau cellulaire et le niveau du cerveau entier



REFERENCES

[1] Jönsson L et al. (2022). The Costs of Dementia in Europe: An Updated Review and Meta-analysis. *Pharmaco Economics*.

[2] Livingston G et al. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The Lancet*.

[3] Alzheimer's Association. 2022 Alzheimer's Disease Facts and Figures. (2022). *Alzheimers Dement*.

Optimiser l'organisation des flux de patients au niveau territorial

Louis DUBOIS

Une approche centrée sur l'adéquation entre des parcours de soins alternatifs et l'offre de soins

Louis Dublois[1], Thierry Garaix[1,2], Yannick Kergosien[3] | [1] LIMOS, UMR 6158, [2] Mines Saint-Étienne, [3] LIFAT, EA 6300, Université de Tours

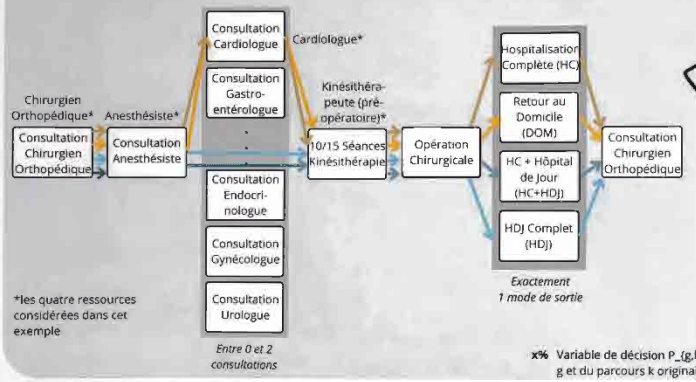
Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0005

Introduction : exemple

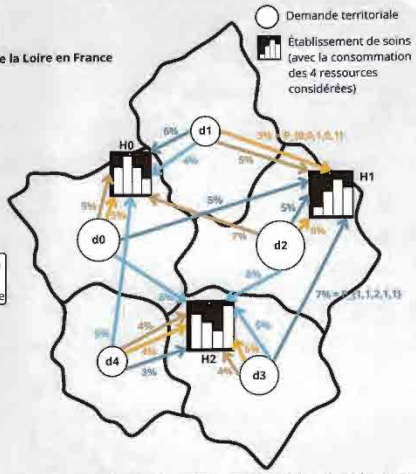
Données issues du SNIIRAM / SNDS sur l'année 2018

- **Sur les procédures chirurgicales de Prothèse Totale de Hanche (PTH) et de Genou (PTG) dans le département de la Loire en France**

Deux groupes de patients :
- PTH avec comorbidité cardiaque : Parcours préférentiel | Parcours alternatif
- PTG sans comorbidité : Parcours préférentiel | Parcours alternatif



*les quatre ressources considérées dans cet exemple



x% Variable de décision P_{g,k,r,a,h} : pourcentage (sur la demande minimum d) de patients du groupe g et du parcours k originaires de la région r qui effectuent leur activité a dans l'établissement h

Modèle d'optimisation des flux de patients

$$\max \ (1-\alpha) \cdot \sum_{g,k} c_{g,k} \cdot d \cdot P_{g,k}$$

$$\max \ \alpha \cdot \sum_{g,k,r,a,h} w_{g,r,h} \cdot d \cdot P_{g,k,r,a,h}$$

$$\sum_{k} P_{g,k,r} \geq d_{g,r}$$

$$\sum_{g,k,r,a} d \cdot P_{g,k,r,a,h} \cdot t_{g,k,a,l} \leq m_{h,l} + \Delta^{+}_{h,l} - \Delta^{-}_{h,l}$$

$$Q_{g,k,r,a,h} \geq P_{g,k,r,a,h} - P_{g,k,r,a',h}$$

$$\sum_{g,k,r,a,h} Q_{g,k,r,a,h} \leq f \cdot \sum_{g,k} n_{g,k} \cdot P_{g,k}$$

$$\sum_{h} \Delta^{+}_{h,l} - \sum_{h} \Delta^{-}_{h,l} = 0$$

Critères

- Qualité de soins
- Préférences des patients à se faire soigner dans tel ou tel établissement
- Équité dans l'accès au soins (spatial et clinique)

Caractéristiques

- Parcours de soins alternatifs par groupe
- Contrainte de capacité sur les ressources
- Transferts de patients à chaque étape du parcours
- Transferts de ressources

Données Demande

- Groupes de patients (GHMs, etc.)
- Nombre minimum de patients à traiter
- Répartition territoriale par groupe et région
- Pourcentages min/max par groupe

Données Parcours de Soins

- Parcours de soins définis par groupe
- Niveau de qualité de soins par parcours
- Activités ordonnées par parcours

DONNÉES → MODÈLE D'OPTIMISATION → SOLUTION

Données Établissements

- Localisation
- Présence de spécialités, de ressources spécifiques
- Capacités par type de ressources (unité)

Données Ressources

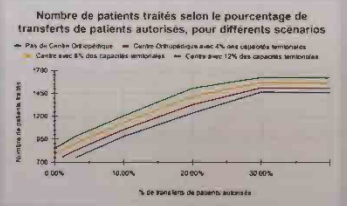
- Type de ressources
- Consommation par activité
- Quantité transférable par type de ressources et par établissement

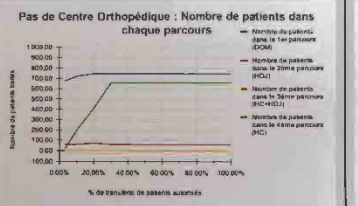
Solution

- Flux de patients par activité et par établissement
- Transferts de patients
- Réallocation de ressources

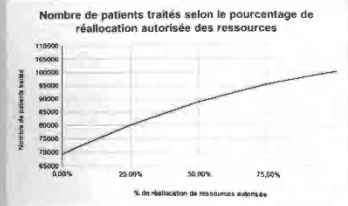
Exemples de résultats

- **Sur les procédures PTH / PTG, et considérant 42 groupes différents (selon les comorbidités) répartis entre PTH et PTG, et avec 4 parcours (DOM, HDJ, HC+HDJ, HC) par groupe**





- **Sur la prestation globale de soins de santé dans la région du Queensland en Australie : cas d'étude issue de la littérature (Burdett et al., 2023)**



Riemannian Flow Matching for Brain Connectivity Matrices Via Pullback Geometry

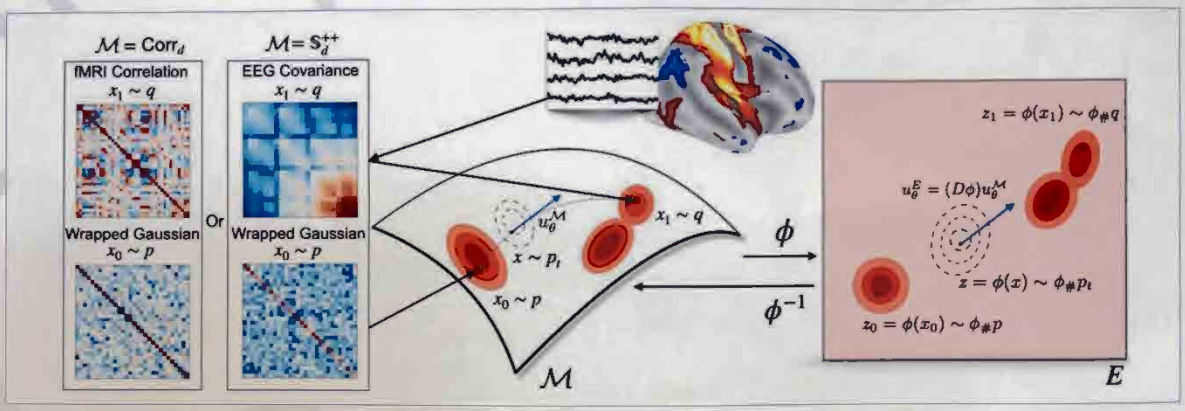
Antoine Collas, Ce Ju, Nicolas Salvy, and Bertrand Thirion
Inria, CEA, and University Paris-Saclay, Palaiseau, France

Generating realistic brain connectivity matrices is **crucial** for:

1. Analyzing population heterogeneity in brain organization;
2. Improving the understanding of neurological disorders;
3. Augmenting datasets in complex classification tasks.

Functional connectivity matrices reside in mathematically constrained spaces, which can be represented as Riemannian manifolds. However, using Riemannian frameworks often involves redefining fundamental operations (e.g., geodesics, norms, and integration), resulting in computationally inefficient generative modeling.

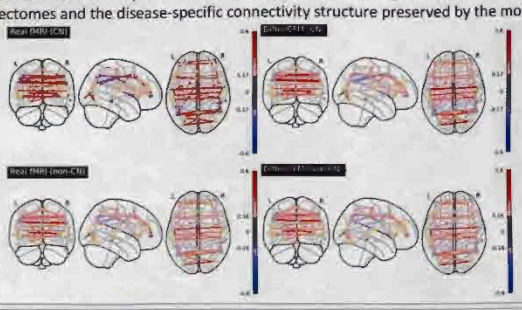
DIFFEOCFM is a deep generative model on matrix manifolds. It reformulates Riemannian conditional flow matching on a pullback manifold as conventional conditional flow matching in Euclidean space, via a global diffeomorphism. This allows both training and sampling to be carried out efficiently in Euclidean space, while remaining equivalent to operating on manifolds. On the left (see the following figure), the fMRI correlation and EEG spatial covariance matrices lie on their own manifolds. These matrices are mapped to Euclidean space through a diffeomorphism. A time-dependent vector field is then trained in Euclidean space, and integration is performed in this space before mapping back via the inverse of the diffeomorphism to yield connectivity manifold-constrained matrices.



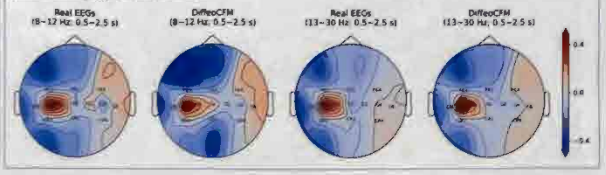
Performance of generative models on 3 fMRI and 2 EEG datasets, evaluated with Quality and Classification Accuracy Score (CAS) metrics. Quality metrics assess alignment with the real distribution. CAS metrics evaluate downstream performance: a classifier is trained on generated data to predict disease status (fMRI) or motor imagery class (EEG), and tested on held-out real samples. REAL DATA uses real samples to compare training and test distributions, serving as an empirical upper bound.

Dataset	Method	Quality Metrics			CAS Metrics	
		α-Precision ↑	β-Recall ↑	α,β-F1 ↑	ROC-AUC ↑	F1 ↑
ABIDE	REAL DATA	0.85 ± 0.04	0.74 ± 0.06	0.79 ± 0.08	0.67 ± 0.05	0.59 ± 0.06
	DIFFEOGAUSS	0.56 ± 0.08	0.28 ± 0.07	0.37 ± 0.11	0.67 ± 0.04	0.52 ± 0.07
	TRIANGDDPM	0.04 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.58 ± 0.04	0.56 ± 0.06
	TRIANGCFM	0.06 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	0.61 ± 0.04	0.42 ± 0.18
	DIFFEOCFM	**0.74 ± 0.05**	**0.46 ± 0.06**	**0.57 ± 0.08**	**0.68 ± 0.03**	**0.60 ± 0.04**
ADNI	REAL DATA	0.88 ± 0.04	0.90 ± 0.04	0.89 ± 0.06	0.63 ± 0.04	0.64 ± 0.04
	DIFFEOGAUSS	0.03 ± 0.03	0.53 ± 0.07	0.06 ± 0.05	0.60 ± 0.06	0.16 ± 0.08
	TRIANGDDPM	0.02 ± 0.01	0.00 ± 0.00	0.00 ± 0.01	0.54 ± 0.04	0.28 ± 0.03
	TRIANGCFM	0.03 ± 0.00	0.00 ± 0.00	0.01 ± 0.01	0.54 ± 0.05	0.26 ± 0.15
	DIFFEOCFM	**0.59 ± 0.15**	**0.77 ± 0.03**	**0.67 ± 0.17**	**0.62 ± 0.04**	**0.35 ± 0.09**
OASIS-3	REAL DATA	0.86 ± 0.06	0.89 ± 0.04	0.87 ± 0.08	0.75 ± 0.04	0.64 ± 0.04
	DIFFEOGAUSS	0.49 ± 0.09	0.30 ± 0.08	0.37 ± 0.12	**0.71 ± 0.07**	0.45 ± 0.05
	TRIANGDDPM	0.04 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	0.52 ± 0.04	0.47 ± 0.06
	TRIANGCFM	0.05 ± 0.01	0.00 ± 0.01	0.01 ± 0.01	0.52 ± 0.08	0.32 ± 0.19
	DIFFEOCFM	**0.64 ± 0.08**	**0.38 ± 0.05**	**0.48 ± 0.09**	**0.71 ± 0.02**	**0.58 ± 0.03**
BNCI 2014-002	REAL DATA	0.69 ± 0.04	0.62 ± 0.05	0.65 ± 0.07	0.83 ± 0.01	0.75 ± 0.02
	DIFFEOGAUSS	0.45 ± 0.04	**0.76 ± 0.01**	0.56 ± 0.05	0.80 ± 0.02	**0.73 ± 0.02**
	TRIANGDDPM	0.43 ± 0.04	0.10 ± 0.02	0.16 ± 0.04	0.54 ± 0.02	0.17 ± 0.12
	TRIANGCFM	0.48 ± 0.03	0.22 ± 0.03	0.30 ± 0.05	0.54 ± 0.03	0.21 ± 0.13
	DIFFEOCFM	**0.63 ± 0.05**	0.63 ± 0.05	**0.63 ± 0.07**	**0.81 ± 0.02**	**0.73 ± 0.02**
BNCI 2015-003	REAL DATA	0.89 ± 0.01	0.89 ± 0.01	0.89 ± 0.01	0.73 ± 0.01	0.67 ± 0.01
	DIFFEOGAUSS	0.84 ± 0.01	**0.89 ± 0.01**	0.87 ± 0.02	**0.73 ± 0.00**	**0.68 ± 0.01**
	TRIANGDDPM	0.72 ± 0.03	0.55 ± 0.02	0.62 ± 0.04	0.59 ± 0.02	0.66 ± 0.02
	TRIANGCFM	0.75 ± 0.08	0.74 ± 0.02	0.74 ± 0.08	0.62 ± 0.03	0.54 ± 0.10
	DIFFEOCFM	**0.92 ± 0.02**	0.86 ± 0.01	**0.89 ± 0.02**	**0.73 ± 0.01**	**0.68 ± 0.01**

Class-conditional fMRI functional connectome plotting using the Fréchet mean (ADNI). Each panel displays class conditional fMRI functional connectomes using the Fréchet mean of correlation matrices computed with respect to the generation diffeomorphism. Left: real data from held-out test subjects; right: samples generated by DIFFEOCFM. The comparison illustrates both the fidelity of generated connectomes and the disease-specific connectivity structure preserved by the model.



Group-level topographic map using the first CSP's spatial filter derived from real EEGs (BNCI2015-001). Generated data by DIFFEOCFM Each map shows the first CSP's spatial filter across 12 subjects in the α (8 – 12 Hz) and β (13 – 30 Hz) frequency bands during the first 2 seconds following stimulus onset. Filters from DIFFEOCFM at the group-level closely resemble real EEGs, preserving discriminative patterns of motor imagery classification.



Funded by ANR 22-PESN-0011 under the France 2030 program, managed by the Agence Nationale de la Recherche (ANR).

REFERENCES
[1] Antoine Collas, Ce Ju, Nicolas Salvy, Bertrand Thirion. "Riemannian Flow Matching for Brain Connectivity Matrices via Pullback Geometry." The Thirty-Ninth Annual Conference on Neural Information Processing Systems (NeurIPS), 2025.

Narrative review on the clinical evaluation of AI-based digital medical devices from a methodological perspective

Sandrine BOULET[1], Vincent THÉVENET[2,3], Judith ABECASSIS[1], Corinne COLLIGNON[5], Bruno GIRAUDEAU[6,7], Edouard LHOMME[2,3,8], François PETIT[9], Raphaël PORCHER[9], Laura RICHERT[2,3,8,10], Florence SAILLOUR[2,3], Moreno URSINO[1], Gaël VAROQUAUX[4], Vincent VERCAMER[11], Sarah ZOHAR[1], Rodolphe THIÉBAUT[2,3,8]

[1] Inserm, UMRS 1346, Université Paris Cité, Inria, HeKA, F-75015 Paris, France; [2] Université de Bordeaux, INSERM, Bordeaux Population Health Research Center, U1219, Bordeaux, F-33000, France; [3] Service d'Information Médicale, CHU de Bordeaux, Bordeaux, F-33000, France; [4] Inria, Soda, Palaiseau, France; [5] Haute Autorité de Santé, Saint Denis, France; [6] Université de Tours, Université de Nantes, INSERM, SPHERE U1246, Tours, France; [7] Inserm CIC1415, CHRU de Tours, Tours, France; [8] Inria, SISTM, Bordeaux, France; [9] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and Statistics (CRESS), F-75004 Paris, France; [10] Université de Bordeaux, INSERM, CHU Bordeaux, Institut Bergonié, CIC 1401, Euclid F CRIN clinical trials platform, F-33000 Bordeaux, France; [11] Digital Health Delegation, French Ministry of Health, Paris, France

CONTEXT

While digital medical devices (DMDs) are often clinically evaluated in a framework similar to that of drugs, they have many specific features of their own, which means that their entire evaluation process needs to be rethought differently from that of drugs.

Objective: To propose a narrative review of these characteristics and the challenges they could imply for the evaluation of DMDs.

MORE SPECIFIC ENDPOINTS

- Importance of organizational impact [1]
- Medical staff adherence and patient compliance: problem of black boxes, fear of data leaks...
- Validity of digital endpoints
- Clinical outcomes may depend on the training (learning curve), competence, and experience of the user.

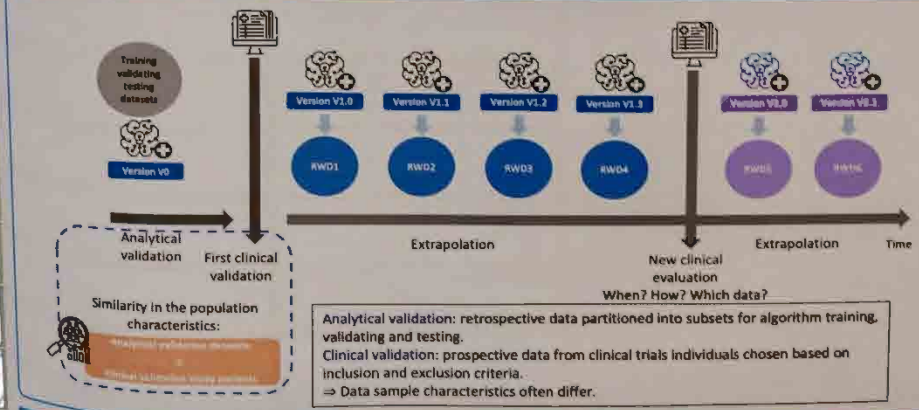
STUDY DESIGN

Randomized Controlled Trials	Real-world studies
+) No selection bias	-) Selection bias
+) High internal validity	-) Time-dependent confounding (individuals may use other interventions)
+) Gold standard for causal inference	-) More missing data
-) Do not consider learning curve	+) Monitor clinical evidence variations over time, by considering the learning curve, and across different users [2]
-) Time-consuming compared to DMD rapid evolution	+) Consider improvement of the device without repeating RCTs
-) Small samples	+) Large and heterogeneous samples
-) Relatively short-term follow-up	+) Long-term follow-up
-) Limited real-world applicability	+) Closer to real clinical practice (higher external validity)
	+) Better economic estimation
-) Expensive	+) Lower costs
-) Measurement or surveillance bias, Hawthorne effect	+) no Hawthorne effect

Comparators:

- Active comparator: Preferred when available.
- Sham: mimics treatment to maintain blinding, +) easier positive results, -) difficult to design a sham that looks like a plausible treatment, -) highlights key intervention components, -) the patient has all the "burden" linked to the device use.
- Standard of care: preferred by industry, +) easier positive results compared to competitor, -) Unethical if an effective comparator exists.
- Control that does not include technology: +) no evolution over time, -) digital placebo effect [3].
- DMD as control: requires the latest version, +) no digital placebo effect, -) may evolve over time.

STUDY POPULATION & WHEN TO EVALUATE?



Analytical validation: retrospective data partitioned into subsets for algorithm training, validating and testing.
Clinical validation: prospective data from clinical trials individuals chosen based on inclusion and exclusion criteria.
⇒ Data sample characteristics often differ.

DMDs' characteristics:

- Fast development
- Often updated, even post-market
- Can be developed using continuous learning systems ⇒ No fixed version.

Questions:

- When to perform the clinical evaluation? With what data?
- What change of the DMD should lead to a new evaluation?

⇒ Need for a new study depends on the extrapolation possibility between two versions [4].

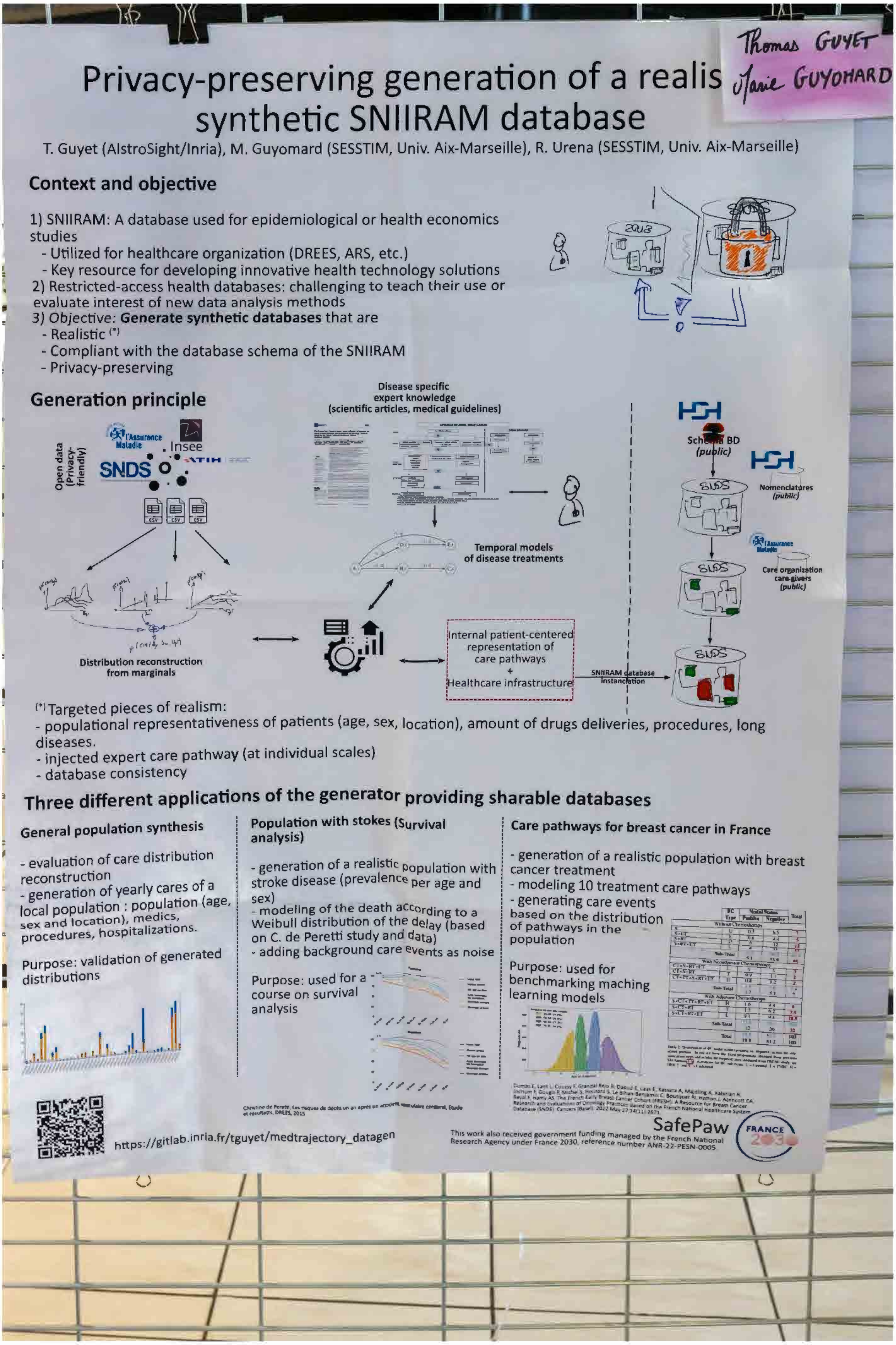
DISCUSSION

- No formal and specific recommendations for clinical trial designs.
- Necessary to develop new, more flexible approaches, such as adaptive trial designs or assessment methods based on real-time data, to take account of the specific features of DMDs.
- Continuous learning implies continuous clinical evaluation and therefore continuous corrections and adaptations.

ACKNOWLEDGMENTS & REFERENCES

[1] HAS, Organisational impact map for health technology assessment (2020).
[2] Tarricone R et al., *Eur Respir Rev*, (2016).
[3] Torous J et al., *Lancet Psychiatry* (2016).
[4] HAS, Evolutions incrémentales des dispositifs médicaux : évaluation par la CNEDiMTS selon leur impact (2021).

This work was partially supported by a French government grant managed by the Agence Nationale de la Recherche under the France 2030 program, reference SMATCH ANR-22-PESN-0003


Thomas GUYET
Marie GUYOMARD
Privacy-preserving generation of a realis[tic] synthetic SNIIRAM database
T. Guyet (AIstroSight/Inria), M. Guyomard (SESSTIM, Univ. Aix-Marseille), R. Urena (SESSTIM, Univ. Aix-Marseille)
Context and objective
1) SNIIRAM: A database used for epidemiological or health economics studies
- Utilized for healthcare organization (DREES, ARS, etc.)
- Key resource for developing innovative health technology solutions
2) Restricted-access health databases: challenging to teach their use or evaluate interest of new data analysis methods
3) Objective: Generate synthetic databases that are
- Realistic (*)
- Compliant with the database schema of the SNIIRAM
- Privacy-preserving
Generation principle
Open data (Privacy-friendly)
l'Assurance Maladie
Insee
SNDS
Disease specific expert knowledge (scientific articles, medical guidelines)
Temporal models of disease treatments
Distribution reconstruction from marginals
Internal patient-centered representation of care pathways + Healthcare infrastructure
SNIIRAM database instanciation
Schema BD (public)
Nomenclatures (public)
Care organization care givers (public)
(*) Targeted pieces of realism:
- populational representativeness of patients (age, sex, location), amount of drugs deliveries, procedures, long diseases.
- injected expert care pathway (at individual scales)
- database consistency
Three different applications of the generator providing sharable databases
General population synthesis
- evaluation of care distribution reconstruction
- generation of yearly cares of a local population : population (age, sex and location), medics, procedures, hospitalizations.
Purpose: validation of generated distributions
Population with stokes (Survival analysis)
- generation of a realistic population with stroke disease (prevalence per age and sex)
- modeling of the death according to a Weibull distribution of the delay (based on C. de Peretti study and data)
- adding background care events as noise
Purpose: used for a course on survival analysis
Care pathways for breast cancer in France
- generation of a realistic population with breast cancer treatment
- modeling 10 treatment care pathways
- generating care events based on the distribution of pathways in the population
Purpose: used for benchmarking maching learning models
https://gitlab.inria.fr/tguyet/medtrajectory_datagen
SafePaw
This work also received government funding managed by the French National Research Agency under France 2030, reference number ANR-22-PESN-0005.
FRANCE 2030

TANAT: A LIBRARY FOR THE ANALYSIS OF TEMPORAL SEQUENCES

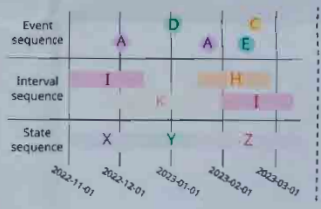
THOMAS GUYET AND ARNAUD DUVERMY, INRIA AISTROSIGHT

TanaT is an **open-source Python framework** for temporal sequence analysis. It handles data made of **timestamped observations described by qualitative and quantitative features**. Such longitudinal data arise in many fields (medicine, social sciences, education, activity traces, ...) and require dedicated methods. ***TanaT*** aims to empower data analysts with a **coherent toolbox for handling temporal sequences at all stages of the analysis process**: data loading, data pre-processing and transformation, data analysis and data visualization.

DATA MODEL: ENTITIES, SEQUENCES AND TRAJECTORIES

TanaT provides a unified model to represent complex temporal data.
- **Entities:** the atomic objects (events, intervals, state).
- **Sequences:** collections of entities with shared features.
- **Trajectories:** groups of sequences with additional static attributes.

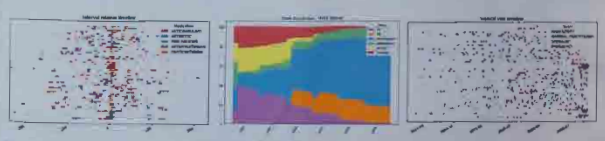
Flexible time representation (datetime or numeric)



CORE FRAMEWORK FUNCTIONALITIES

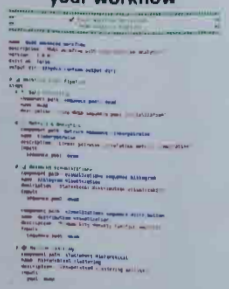
Synthetic data generation for statistical power analysis and comprehensive benchmarking of algorithms

Controllable simulation with customizable:
- vocabulary,
- generation methods,
- time sampling strategy

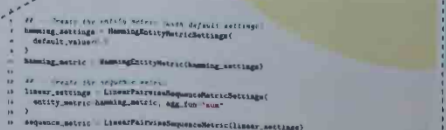


Customizable visualizations for both **individual** and **pooled sequences**

Manage and customize your workflow



- **YAML-based description** of complete workflows
- Easy bash execution for **reproducibility and automation**



Compose a trajectory metric **by combining entity, sequence and trajectory metrics** (including time-warping support).

Data Integration & Preprocessing

Import from **various sources**:
- pandas
- CSV
- databases
- parquet file

Data wrangling:
- pattern-based selection,
- query filtering,
- (static) feature engineering

Export to **other data structure**:
- tensors,
- tabular data,
- time series

TensorFlow PyTorch aeon

Integrated survival analysis for all sequence types

Sksurv modules trainable in just a few lines.

TANAT
TEMPORAL ANALYSIS OF TRAJECTORIES
https://gitlab.inria.fr/tanat/core/tanat

This work has been partly funded by Chaire AIRacles (Inria/APHP) this work also received government funding managed by the French National Research Agency under France 2030, reference number ANR-22-PESN-0005.

Modeling the cortical fold opening process during aging

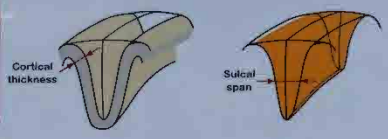
Théo Delmaire[1], Wenqi Shu-Quartier[1], Vincent Frouin[1], Miguel Guevara[1], Clara Fischer[1], Jean-François Mangin[1,2]

1 Université Paris-Saclay, CEA, CNRS, Neurospin, UMR9027 Baobab, Gif-sur-Yvette, France
2 CATI, UNSZ-UAR2031, CEA, ICM, Sorbonne Université, CNRS, INSERM, APHP, Île de France, France
Corresponding author: theo.delmaire@cea.fr

théo DELMAIRE

Context

The patterns of cortical **atrophy associated** with the spread of neurodegenerative diseases and **the effects of aging** are generally studied using features like cortical thickness, surface area and gyrification index [1]. Research on **the gradual opening of cortical folds** due to the constant reduction of tissue is rare [2-3]. One study has shown that **sulcal opening follows the same patterns as cortical thickness**, but earlier. However, to our knowledge, no study has examined this phenomenon at **the individual level**. Our work aims to understand these **morphological alterations** and link them to biological markers such as APOE 4 genes.



Materials and Methods

1. Observing an **atrophy acceleration** on a general population (UK Biobank)
2. Comparison between **transversal** (UK Biobank) and **longitudinal** cohorts (SENIOR)
3. observing the **longitudinal trajectories** of the opening of the sulci on a longitudinal cohort (SENIOR)

Data

- **Two cohorts :**
  1. Transversal cohort : **UK Biobank**
  2. Longitudinal cohort: **SENIOR**
     - **T1-weighted MRIs**
     - **Brain features** (sulcal opening) extracted for each sulcus using **BrainVisa**

UK Biobank	SENIOR
• 20,541 subjects	• 142 subjects
• Average age 62 years	• Average age 60 years
	• One acquisition per year during 10 years

1. **Piecewise linear model**, then a **bootstrap** with 100 iterations. 90th percentile observations.
2. **Piecewise linear model**, then a **bootstrap** with 100 iterations on the SENIOR data. **Each acquisition per subject** was considered as an **independent value**, and the **average per year** was calculated. Comparison of **SENIOR** results with those obtained from **UK Biobank** with the 90th percentile.
3. **Extraction** of morphometric data of **sulcal opening** using the **Morphologist** software on the **SENIOR** dataset, and observation of the **evolution of trajectories** over the years.

Results

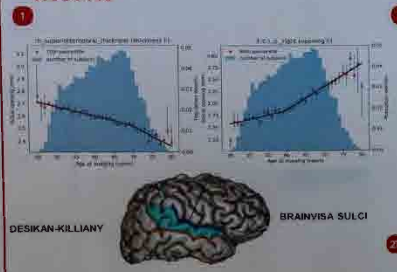


Fig1. Comparison between the sulcal opening of the two sulci around the superior temporal gyrus and the thickness of this gyrus during aging

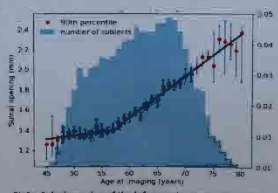


Fig2a. Sulcal opening of the left superior temporal sulcus according to age in women among 20,541 subjects from UKBiobank on the 90th percentile

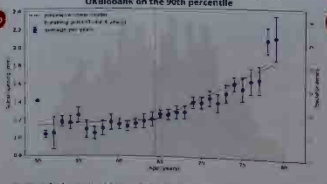


Fig2b. Sulcal opening of the left Superior temporal sulcus according to age in 42 women from SENIOR

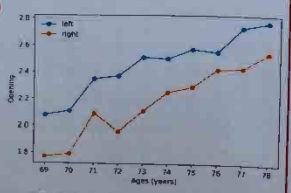


Fig3a. Evolution of the sulcal opening of left and right superior temporal sulci over 10 years for 1 woman SENIOR subject

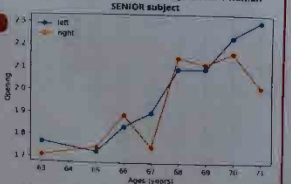


Fig3b. Evolution of the sulcal opening of left and right superior temporal sulci over 10 years for 1 woman SENIOR subject

Discussion

- **Sulcal opening** allows for **earlier** detection of **model brain atrophy** during the aging process than cortical thickness. The **inflection point** occurs at **57 years old** in **women** for **sulcal opening**, while it occurs at **71 years old** for **cortical thickness**.
- A comparison of **sulcal opening** between the **UK Biobank** (90th percentile) and the **SENIOR** cohort (subjects without pathology) shows that sulcal opening is clearly observable in all individuals, even in the absence of pathological risk. These results suggest that this is a **normal physiological process of brain aging**. However, these observations remain indicative: the dynamics of opening vary from one individual to another, and the slope of this evolution may occur earlier or later depending on individual trajectories.
- At the individual level, the phenomenon of sulci opening's acceleration can be observed but depends on the age at which subjects are included in the cohort, as the inflection point may occur before inclusion.

[1] Fjell, A. M (2014). "What is normal in normal aging ? Effects of aging, amyloid and Alzheimer's disease on the cerebral cortex and the hippocampus"
[2] Bertoux, M. (2018). " Sulcal morphology in Alzheimer's disease: an effective marker of diagnosis and cognition"
[3] Tang, H. (2021). " A slower rate of sulcal widening in the brains of the nondemented oldest old "

This work has benefited from funding managed by the National Research Agency under the reference ANR-22-PESN-0010- StratifyAging (France 2030)

Implementation of a Disease Progression Model accounting for covariates

G.CASIMIRO[1], S.KAISARIDI[1], S. TEZENAS DU MONTCEL[1]

1 - Inria, centre of Paris, ARAMIS team, Sorbonne Université, Institut du Cerveau / Paris Brain Institute AP-HP, INSERM, CNRS, University Hospital, Paris, France

Projet financé dans le cadre de France 2030 par l'ANR sous la référence ANR-22-PESN-0016

leaspy https://github.com/aramis-lab/leaspy gabrielle.casimiro@icm-institute.org http://www.aramislab.fr/

Objectives

Disease progression models are promising tools for analyzing longitudinal data presenting multiple modalities. Such models can be used to estimate long-term disease progression and reconstruct individual trajectories. **Inter-patient variability** is often modeled as random perturbations around a fixed reference. However, much of this variability is driven by external factors such as genetic mutations, gender, level of education or socio-economic status.

In this work, we extend the **Disease Course Mapping Model**, a non-linear mixed-effects disease progression model implemented in the opensource library **Leaspy**, implementing a multivariate logistic framework **to explicitly account for covariates.**

Context

Disease Course Mapping model

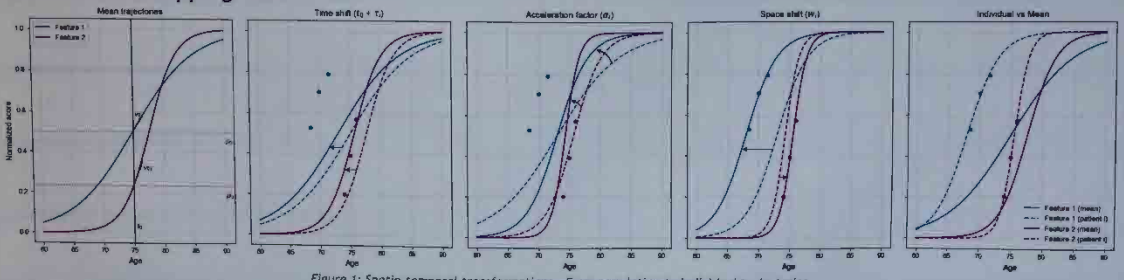


Figure 1: Spatio-temporal transformations - From population to individual trajectories

Population parameters (t_0, p_0, v_0) and **Individual parameters** (τ_i, α_i, w_i)

=

Latent Variables drawn from Prior Distributions

$$\tau_i \sim \mathcal{N}(\bar{\tau}\,;\,\sigma_\tau^2)$$

that depend on Model Parameters to be estimated, and fixed Hyperparameters.

MCMC SAEM Algorithm

1. Sample Latent Variables given posterior distribution
2. Update sufficient statistics by averaging the new samples with previous estimates
3. Update Model Parameters by maximizing the expected complete-data log-likelihood

Methods

Covariate-augmented model

Population Parameters (t_0, p_0, v_0) are now estimated as a deterministic function of the covariates:

$$(t_0, p_0, v_0) = f_\varphi(c)$$

Link Function $f_\varphi(c) = \varphi_{mod} \cdot c + \varphi_{ref}$

New Latent Variables with the following **Prior Distribution:**

$$\begin{bmatrix}\varphi_{mod} \\ \varphi_{ref}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\varphi_{mod} \\ \varphi_{ref}\end{bmatrix}, \begin{bmatrix}\sigma_{mod}^2 & \rho\sigma_{mod}\sigma_{ref} \\ \rho\sigma_{mod}\sigma_{ref} & \sigma_{ref}^2\end{bmatrix}\right)$$

introducing new Model Parameters to be estimated via the **MCMC SAEM Algorithm.**

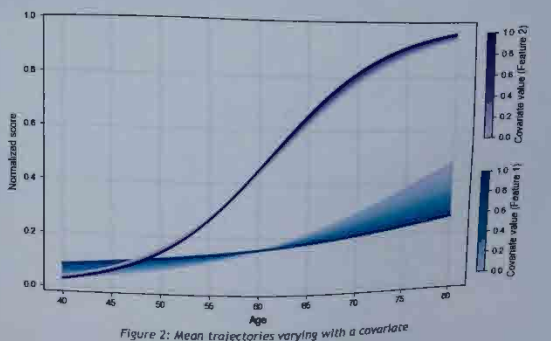


Figure 2: Mean trajectories varying with a covariate

Conclusion

This ongoing work presents a modeling approach to study how a covariate may influence the progression of multiple features over time. The approach is under implementation and will be tested on real data.

References :

- Jean-Baptiste Schiratti, Stéphanie Allassonnière, Olivier Colliot, and Stanley Durrleman. "A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations". In: Journal of Machine Learning Research 18.133(2017), pp.1-33.
- Nemo Fournier and Stanley Durrleman. "A Multimodal Disease Progression Model for Genetic Associations with Disease Dynamics". In: Medical Image Computing and Computer Assisted Intervention (MICCAI). https://hal.archives-ouvertes.fr/hal-04295080.Vancouver,Canada,Oct.2023,pp.601-610.doi:10.1007/978-3-031-43904-9_58.

On the Relevance of ECG Features for Long-Term Survival Prediction - Application to Septic Shock

R. Chalard[1], A. Kachenoura[1], G. Carrault[1], D. Ge[1], A. Kimmoun[2], A. Mebazaa[3], A. Karfoul[1]

1 Univ. Rennes, Inserm, LTSI-UMR 1099, 35000 Rennes, France
2 Faculté de Médecine, Inserm DCAC-UMR 1116, 54511 Vandoeuvre-les-Nancy, France
3 Univ. Paris cité, Inserm, MASCOT-UMR-S 942, 75010, Paris, France
raphael.chalard@univ-rennes.fr

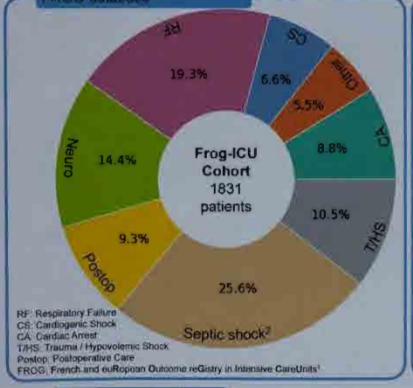
Context & Objective

Context:
- Long-term outcomes of Intensive Care Unit (ICU) patients are poorly documented and often understudied
- Despite being systematically recorded, ECGs are largely underutilized in risk stratification and outcome prediction

Objective:
- Demonstrate that ECG features contain valuable long-term prognostic information about ICU patients survival at one year after the first recorded ECG

Data cohort & Pipeline

FROG database[1]

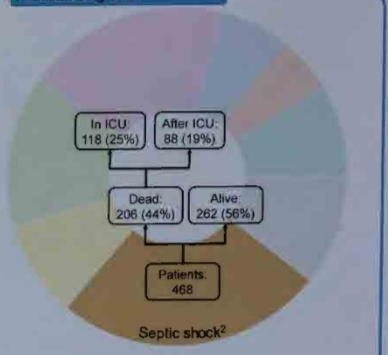


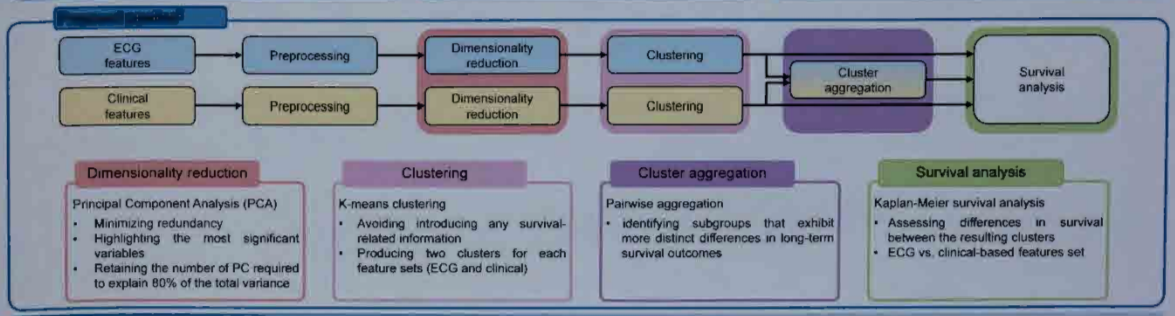
RF: Respiratory Failure
CS: Cardiogenic Shock
CA: Cardiac Arrest
TMS: Trauma / Hypovolemic Shock
Postop: Postoperative Care
FROG: French and euROpean Outcome reGistry in Intensive CareUnits[1]

Features

65 Clinical features		165 ECG features	
Domain	nbr	Domain	nbr
Demographics	4	Electrical axis	2
Severity scores	2	Segment durations	7
Comorbidities	37	Waves amplitudes	7 x 12(leads)
Chronic treatments	22	Waves duration	6 x 12(leads)

Focused diagnosis





Dimensionality reduction

Principal Component Analysis (PCA)
- Minimizing redundancy
- Highlighting the most significant variables
- Retaining the number of PC required to explain 80% of the total variance

Clustering

K-means clustering
- Avoiding introducing any survival-related information
- Producing two clusters for each feature sets (ECG and clinical)

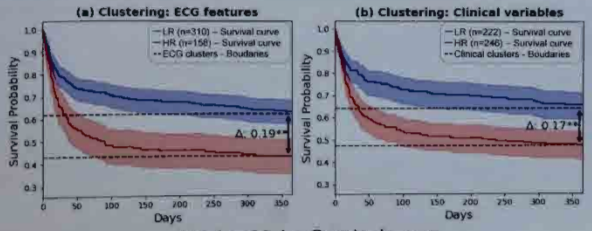
Cluster aggregation

Pairwise aggregation
- identifying subgroups that exhibit more distinct differences in long-term survival outcomes

Survival analysis

Kaplan-Meier survival analysis
- Assessing differences in survival between the resulting clusters
- ECG vs. clinical-based features set

Results



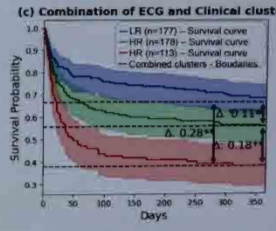
Kaplan-Meier Survival curve

Pairwise aggregation	ECG_{LR}	ECG_{HR}
$Clinic_{LR}$	177	45
$Clinic_{HR}$	133	113

LR: Low-Risk population
MR: Medium-Risk population
HR: High-Risk population
Δ indicates the survival gap. *p < 0.05, **p < 0.005

Cluster aggregation

Two similar-risk groups merged into one group (green)



Conclusions& Perspectives

Conclusion:
- A POC study demonstrated that ECG features (from the first ICU ECG) are as relevant as clinical ones for predicting long-term survival in ICU septic shock patients
- Clusters aggregation highlights some uncertainty in the one-year survival prediction of certain ICU septic shock patients (green cluster)

Perspectives:
- To extend current work to other FROG subpopulations (e.g., Respiratory Failure and Neuro)
- To copp with the uncertainty issue through advanced unsupervised/supervised approaches to enhance one-year survival prediction

References

1. Mebazaa A, et al. Post-ICU discharge and outcome: Rationale and methods of The French and euROpean Outcome reGistry in Intensive Care Units (FROG-ICU) observational study. BMC anesthesiology 2015;15:143.
2. Angus DC, van der Poll T. Severe Sepsis and Septic Shock. New England Journal of Medicine 2013;369:840–851.
3. **Chalard R**, et al. On the Relevance of ECG Features for Long-Term Survival Prediction - Application to Septic Shock. Computing in Cardiology (CinC São Paulo 2025)

Acknowledgments

PEPR Santé Numérique, Projet DIIP-HEART ANR 22-PESN-0018, France