Nantes Université

PLACE DES DONNÉES ET COLLECTIONS BIOLOGIQUES DANS LES ALGORITHMES DE MÉDECINE DE PRÉCISION

JOURNÉE NATIONALE FRBIONET Le 24 septembre 2026

De la valorisation du réseau OFSEP vers un modèle de médecine de précision dans la sclérose en plaques : Projet RHU PRIMUS

Pr Pierre-Antoine Gourraud

Professeur des Universités & Praticien-Hospitalier

CHU Nantes, PHU 11 : Clinique des données, INSERM, CIC 1413, Nantes Université, INSERM, CR2TI

pierre-antoine.gourraud@univ-nantes.fr

CC BY NC

FrBioNet FRANCE BIOBANK NETWORK — FRANCE 2030

Nantes Université

Pôle Santé UFR Médecine — DELPHI DEep Learning Proposal for Healthcare & Innovation NEXT Nantes Université AI & Health data — FRANCE 2030 — anr — Région Pays de la Loire — Nantes Métropole — Nantes Université — Nantes Université — Inserm Institut national de la santé et de la recherche médicale

Nantes Université

- **Conflits d'intérêt potentiels:**

PA Gourraud est le fondateur (2008) de Methodomics. Sa spin-off est Big data Santé (2018- nom commercial «Octopize»). Il est consultant et/ou intervenant pour de grandes entreprises pharmaceutiques ou de dispositifs médicaux. Ses activités sont toutes traitées par une contractualisation universitaire ou hospitalière*(AstraZeneca, Amgen, Biogen, BMS Boston Scientific, Cemka, Cepton, Cook, Docaposte/Heva, Edimark, Ellipses, Elsevier, Grunenthal, Hemovis, Janssen, IAGE, Lek, Lifen, Methodomics, Merck, Mérieux, Novartis, Octopize, Sanofi-Genzyme, TuneInsight)* . PA Gourraud est administrateur bénévole des mutuelles d'assurances AXA (2021). Il n'a aucune activité de prescription de médicaments ou dispositifs médicaux. Il ne perçoit pas de rémunération complémentaire

- ***Potential COI :***

PA Gourraud is the founder of Methodomics (2008) and its spin-off Big data Santé (2018- "Octopize" brand). He consults for major pharmaceutical companies, and start-ups, all of which are handled through academic pipelines (AstraZeneca, Amgen, Biogen, BMS, Boston Scientific, Cemka, Cepton, Cook, Docaposte/Heva, Edimark, Ellipses, Elsevier, Grunenthal, Hemovis, Janssen, IAGE, Lek, Lifen, Methodomics, Merck, Mérieux, Novartis, Octopize, Sanofi-Genzyme, Lifen, TuneInsight). PA Gourraud is a volunteer board member at AXA not-for-profit mutual insurance company (2021). He has no prescription activity with either drugs or devices. He receives no wages from these activities.

Nantes Université

Utilisation secondaire de données de santé par l'exemple

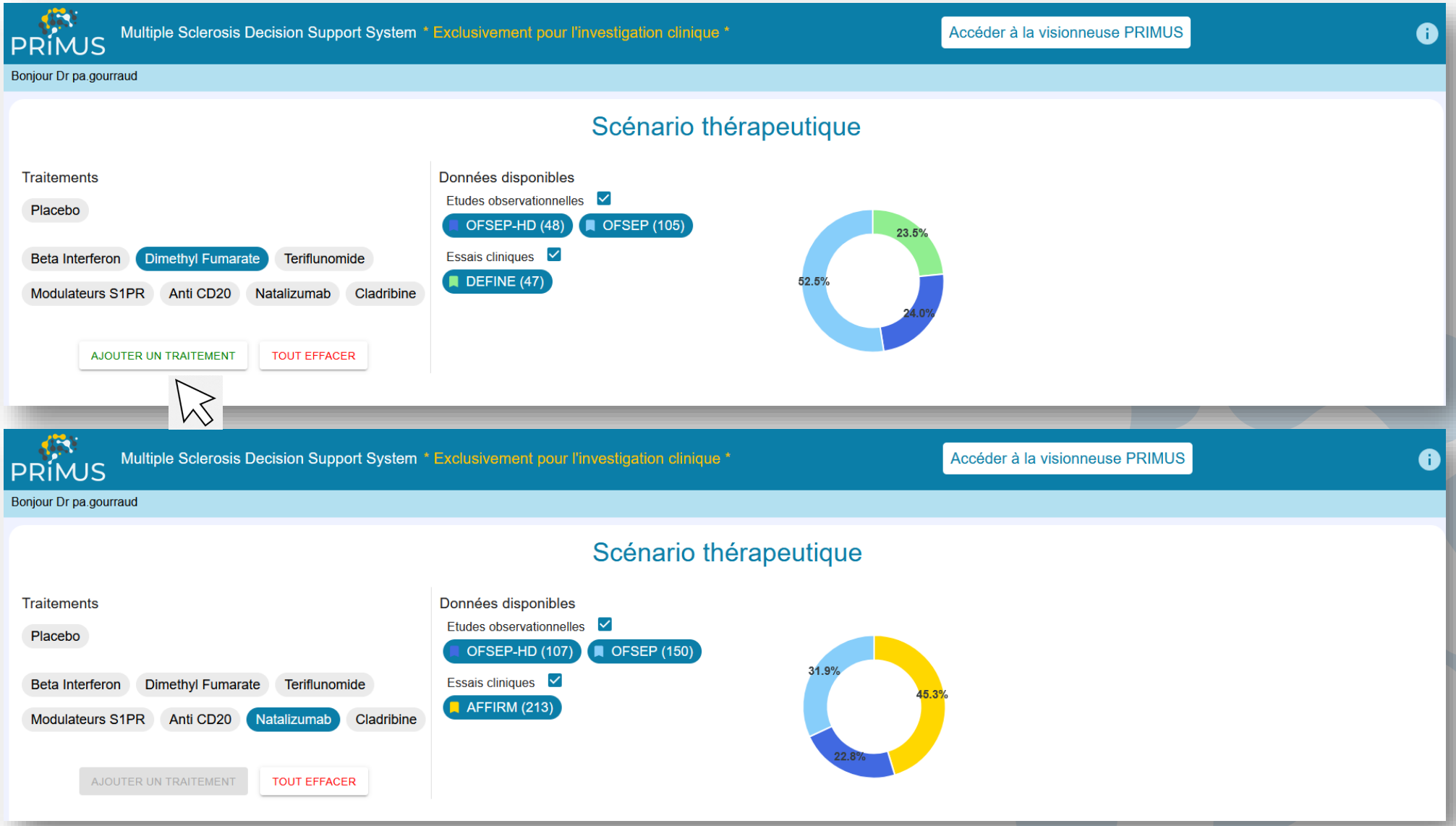
Réseau OFSEP et plus dans le projet RHU PRIMUS

Démonstration - log in – POI// Les données du patients comme point de depart.

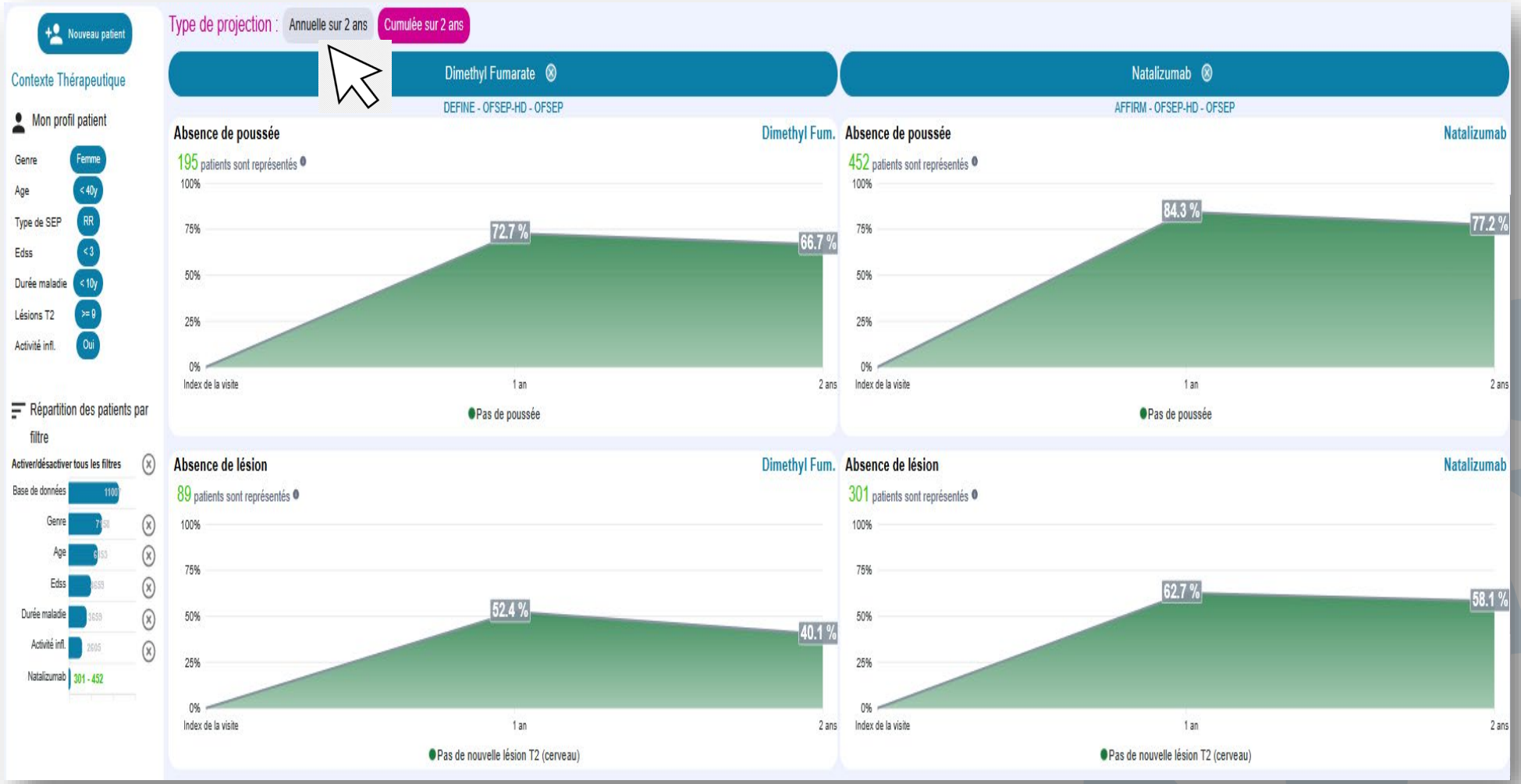
https://primus-cdss.pixyl.io/



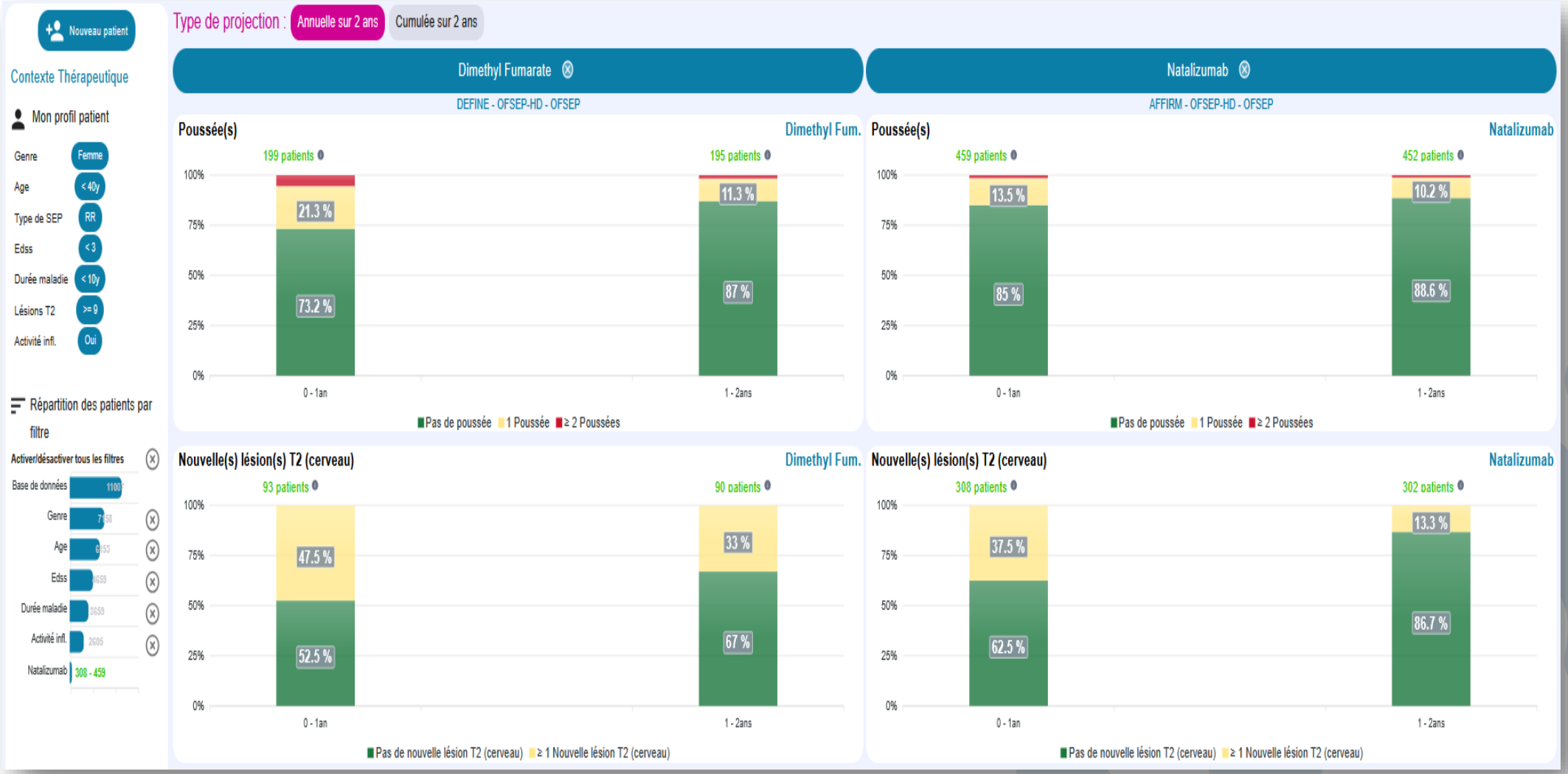
Accès à des données de référence : Observationnelles & RCTs



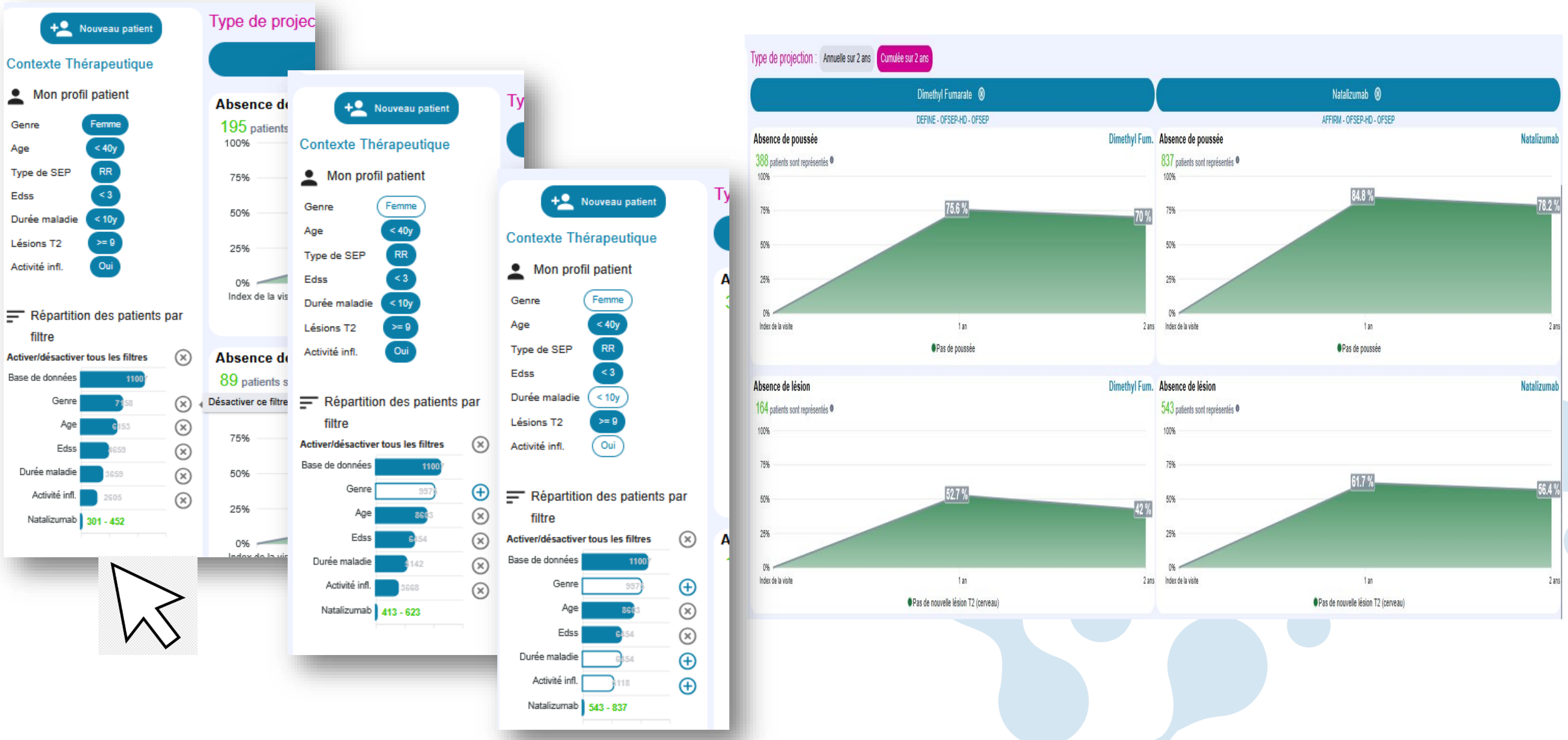
Projection à 2 ans – Patients “similaires”



Projection à 2 ans – Patients “similaires”



Interface collaborative



Sous le capot : 1/3
Des données multi-sources interopérables mises en qualité: un travail

// It is all about DATA + Architecture IT adaptée à la responsabilité initiale de traitement

Randomized Phase III Clinical Trials

JUVISE pharmaceuticals

OPTIMUM
-Ponesimod (n = 471)
-Teriflunomide (n = 473)

MERCK

REFLEX
-No treatment (n = 169)
-Beta Interferon (n = 173)

ORACLE
-No treatment (n = 203)
-Beta Interferon (n = 206)

CLARITY
-No treatment (n = 434)
-Cladribine (n = 430)

CLARITY ext
-No treatment (n = 244)
-Cladribine (n = 98)

Biogen

CONFIRM
-No treatment (n = 363)
-Dimethyl Fum. (n = 359)
-Glatiramer Acet. (n = 350)

DEFINE
-No treatment (n = 408)
-Dimethyl Fum. (n = 410)

AFFIRM
-No treatment (n = 312)
-Natalizumab (n = 627)

SENTINEL
-No treatment (n = 595)
-Beta Interferon + Natalizumab (excluded)

ADVANCE
-No treatment (n = 490)
-Beta Interferon (n = 505)

ASCEND
-No treatment (n = 449)
-Natalizumab (n = 439)

ATTAIN
-Beta Interferon (n = 547)

Observational Studies

OFSEP
Observatoire Français de la Sclérose en Plaques

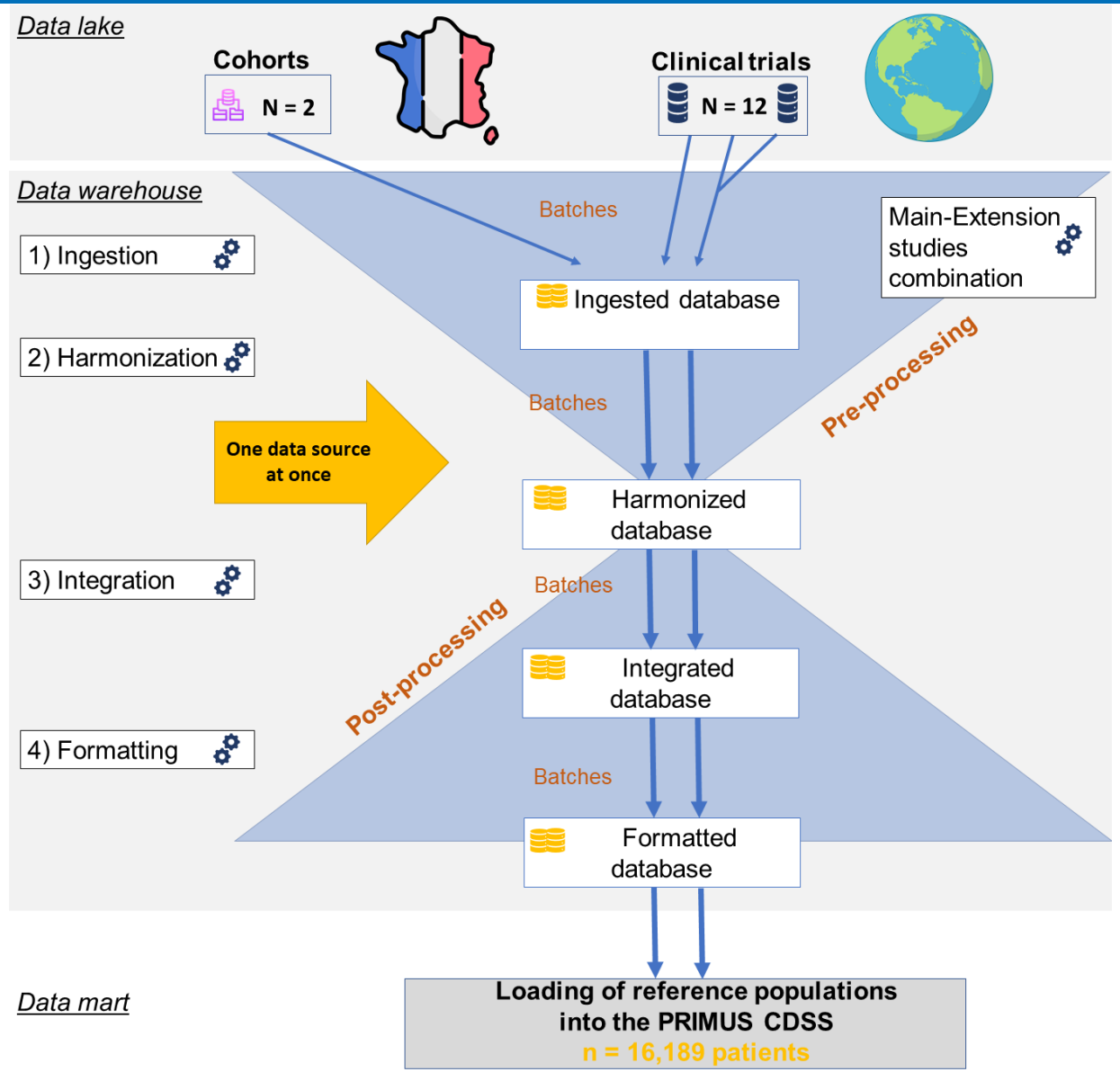
Hight definition cohort (OFSEP-HD)
-All treatments (n = 2 842)

Part of general cohort (OFSEP-Mother)
-All treatments (n = 15 962)

Ker Primus Data Warehouse (BKP)

N = 25 737 Patients Of Reference

// OBVIOUSLY : Data management



“ The Value is in the data transformation more than the necessary yet not usable raw data”

ETL pipeline- *ClinSci*

https://gitlab.com/stanislas.demuth/clinsci

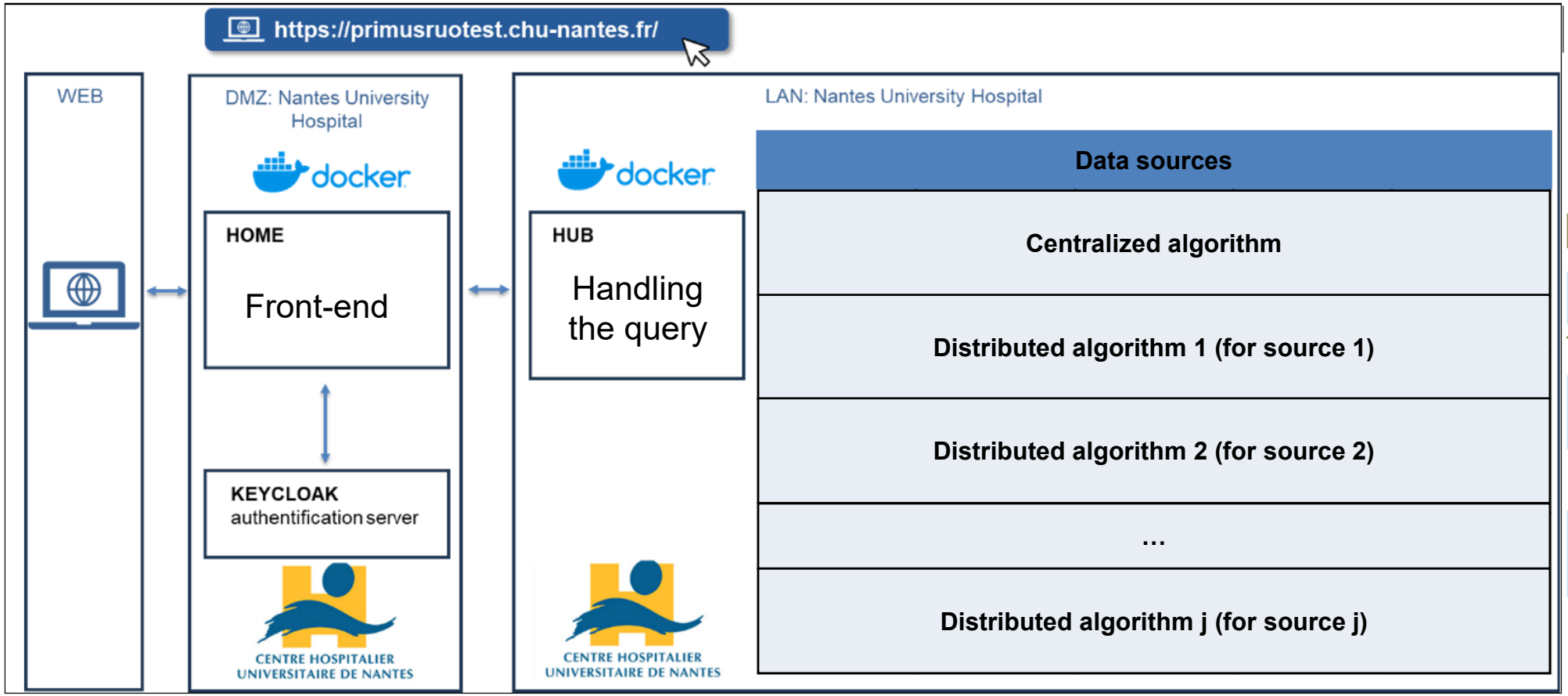
Demuth, Faddeenkov et al. (2024); Are we in a Big Data era for multiple sclerosis? Lessons from integrating clinical trials and observational studies data into the PRIMUS precision medicine platform. DOI: 10.1101/2024.10.17.24315655 (pre-print)

ETL: Extract-Transform-Load

Sous le capot : 2/3

Une infrastructure Fédérée pour garantir le contrôle des usages

// INNOVATION : Distributed infrastructure

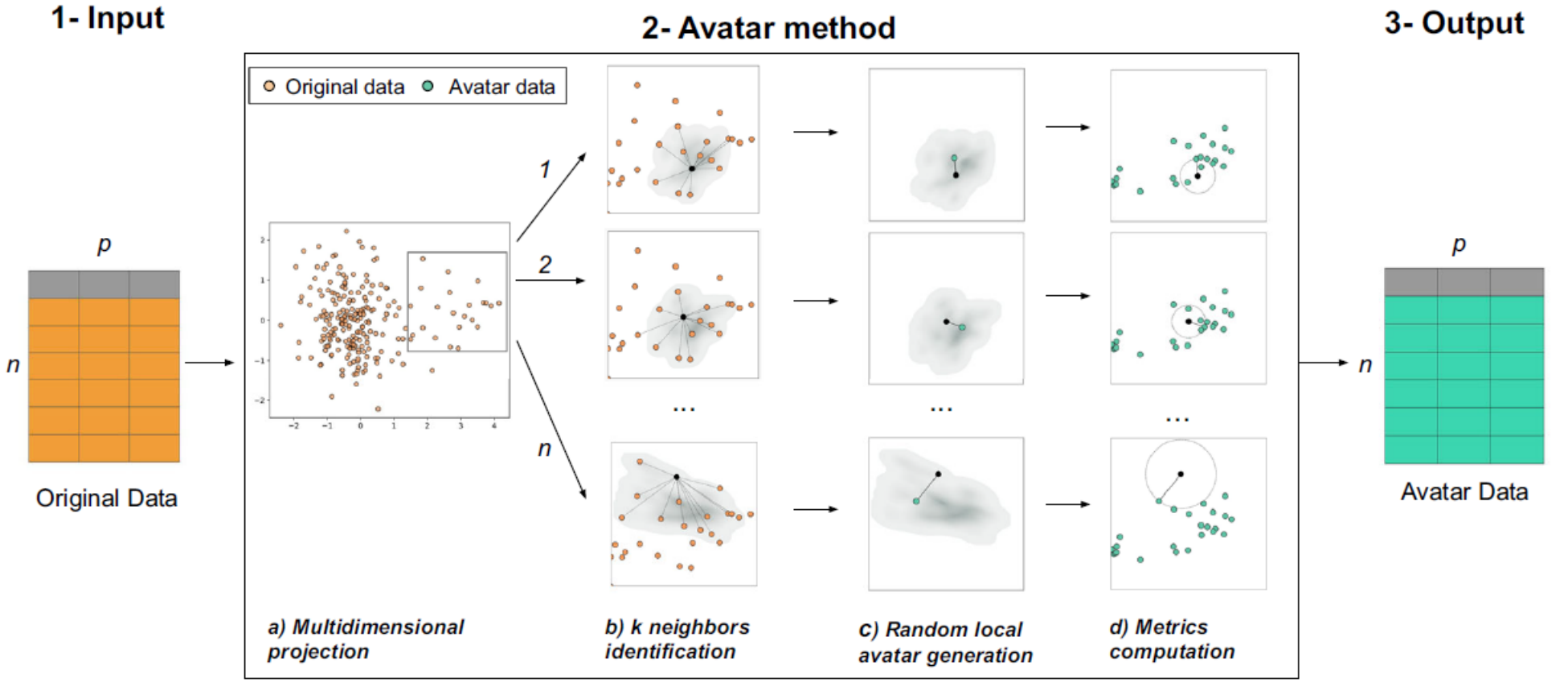


Demuth, Faddeenkov et al. (2024); A collaborative algorithm approach to personalized therapeutic management in multiple sclerosis: a validation study of the PRIMUS clinical decision support system (in preparation)

Sous le capot : 3/3

Des données synthétiques anonymes augmentées « AI ready »

// INNOVATION 2 : AUGMENTED SYNTHETIC ANONYMOUS DATA

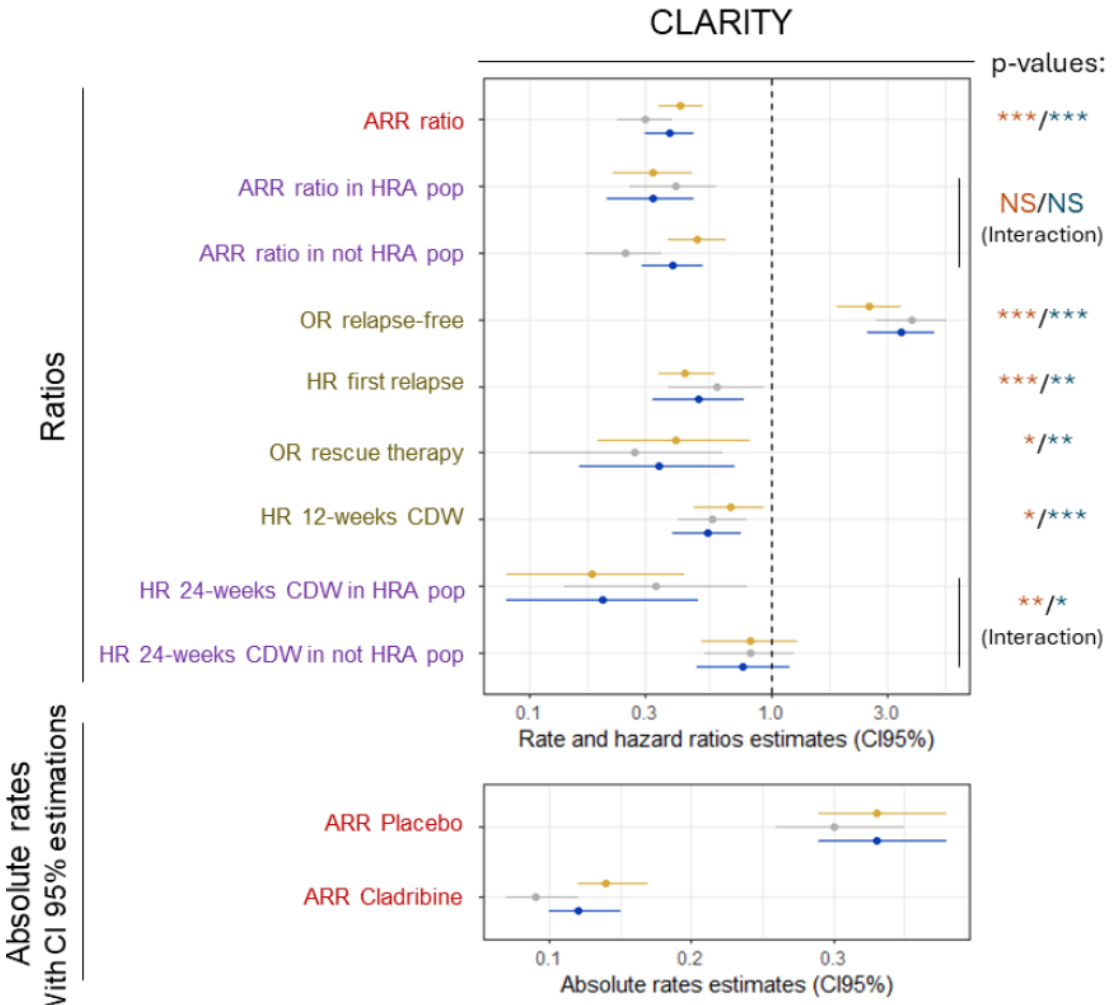


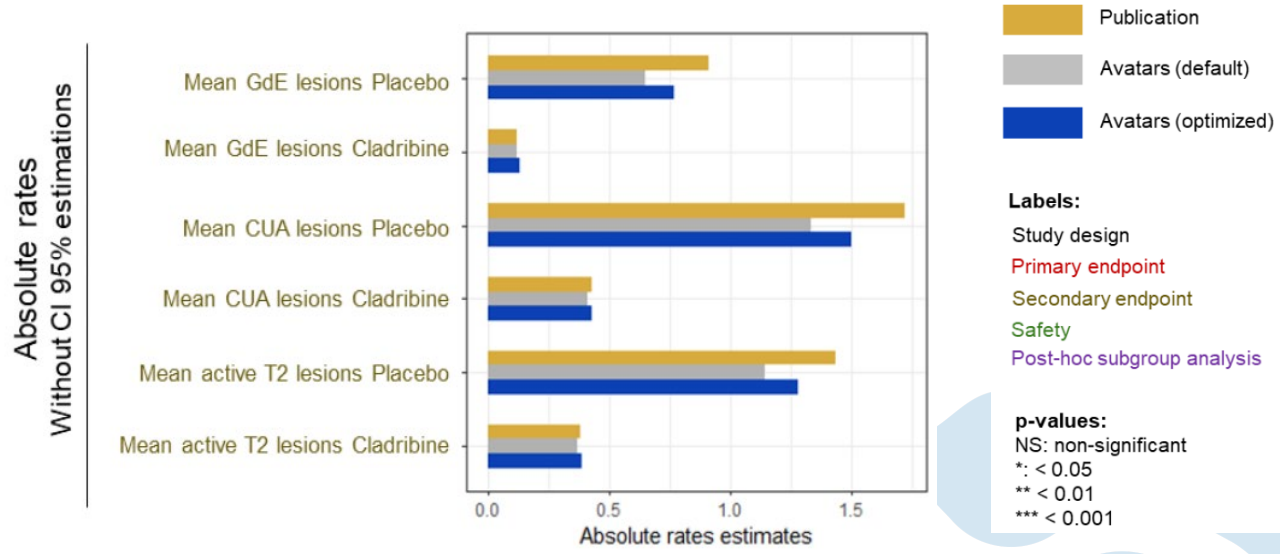
Preservation of representability + ensured privacy according to GDPR due to 1-to-1 correspondence of patient and his avatar

GDPR: EU General Data Protection Regulation

Guillaudeux, Gourraud et al. (2023) "Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis". NPJ Dig. Med. DOI: 10.1038/s41746-023-00771-5

// A TRADE OFF BETWEEN UTILITY And PRIVACY ?





Singling out ✓

Linkability ✓

Inference ✓

Validation both of the ETL pipeline AND the synthetic data

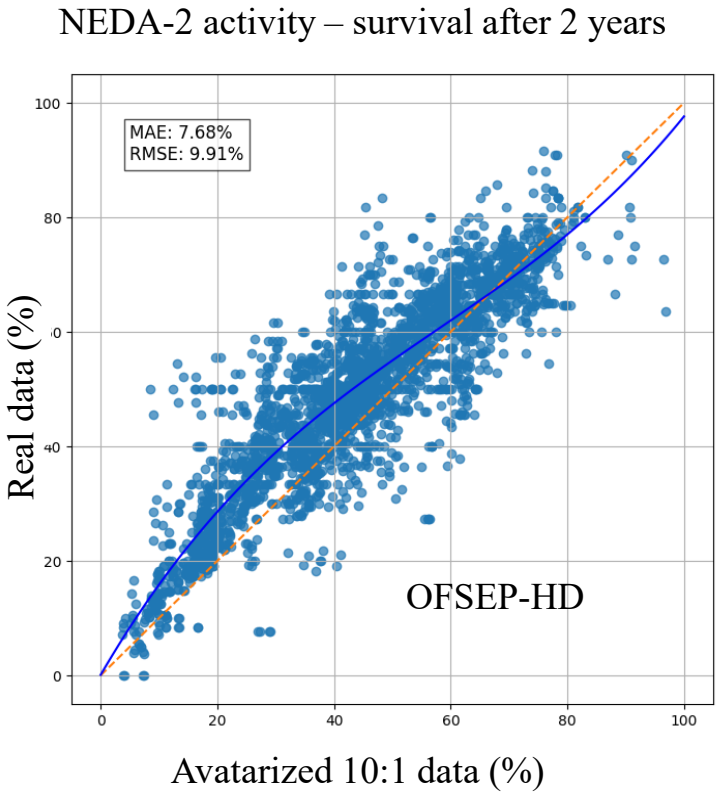
GDPR: EU General Data Protection Regulation

Demuth, Faddeenkov et al. (2025); Privacy-By-Design Generation of Two Virtual Clinical Trials in Multiple Sclerosis and their Release as Open Datasets: Evaluation Study. JMIR. (in process of publication)

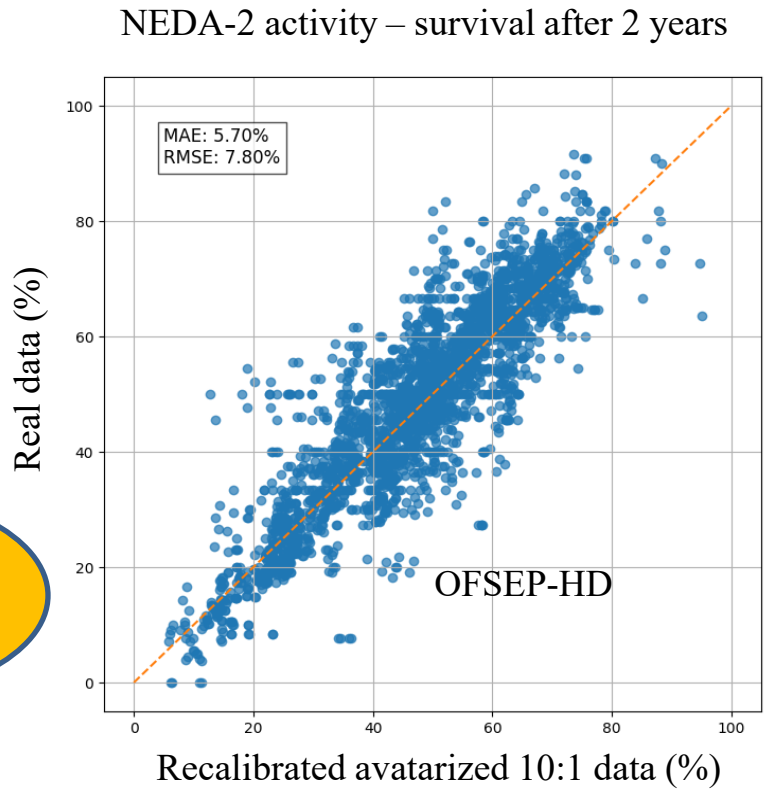
// AUGMENTATION AND RECALIBRATION



Avatari-zation X10



Recalibration with the polynomial



Comparison of all possible queries and polynomial fit

// GDPR COMPLIANCE : privacy comes first !

CNIL. | PROTÉGER les données personnelles
ACCOMPAGNER l'innovation
PRÉSERVER les libertés individuelles

- Hidden rate:
  represents the percentage of individuals in the original dataset whose avatar simulation is not the most similar to them. This metric evaluates the probability of an attack being wrong when it associates an avatar simulation with the individual to whom the avatar simulation is most similar
- Local cloaking:
  Number of avatar simulations that are more similar (i.e. closer in the multidimensional space) to the original data than the one avatar produced from the data

= Nous ne manipulons plus des Données Personnelles

Anonymization criteria	Metric	Definition	Software Editor recommendation (Indicative)	CLARITY (Optimized parameters)
Singling out	Distance to the closest	Median distance between each synthetic data point and their closest reference data point	>0.2	0.31
	Distance to the closest ratio	Median of the ratio of distances between each synthetic data point and their closest and second closet reference data point	>0.3	0.81
Linkability	Column direct match protection	Minimum probability that a variable could be used as a direct identifier	>50%	84.8%
	Row direct match protection	Percentage of synthetic data points that are identical to reference data points	>90%	100%
Inference	Median local cloaking	Median number of avatars more similar to the reference data point of a patient than its own avatar	>5	3
	Hidden rate	Probability of erroneous distance-based matching	>90%	85.0%
	Categorical hidden rate	Probability of erroneous distance-based matching based on categorical variables only	>90%	98.4%

Nantes Université

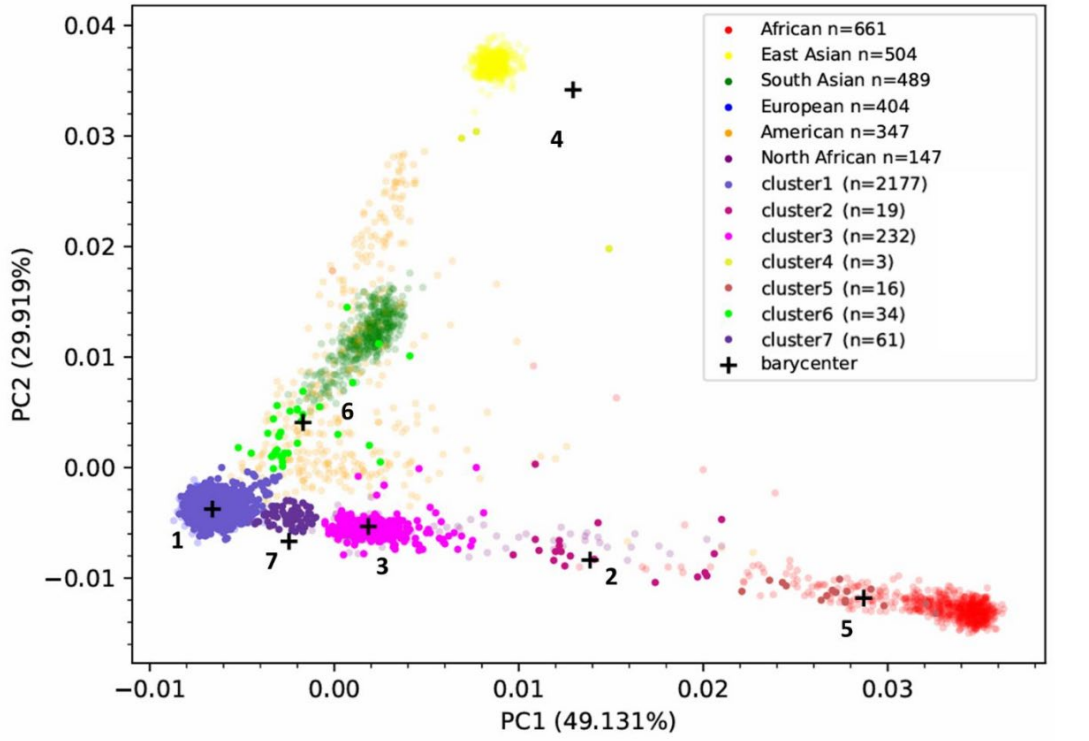
La Biocollection Génétique OFSEP-HD – Sclérose en Plaques

OFSEP-HD Genetics Project : First publication of the French cohort profile

Paris et al. preprint:
https://www.medrxiv.org/content/10.1101/2025.04.28.25326562v2
Github resources : https://github.com/jp3142/

Take home : Genetic Ancestry in 7 clusters

- 85.6% European n= 2177
- Cluster 3 :
  - N=232 North-African MS Patients
  - 51.7% of North-African ancestry patients did not self-identify as such
- Others clusters

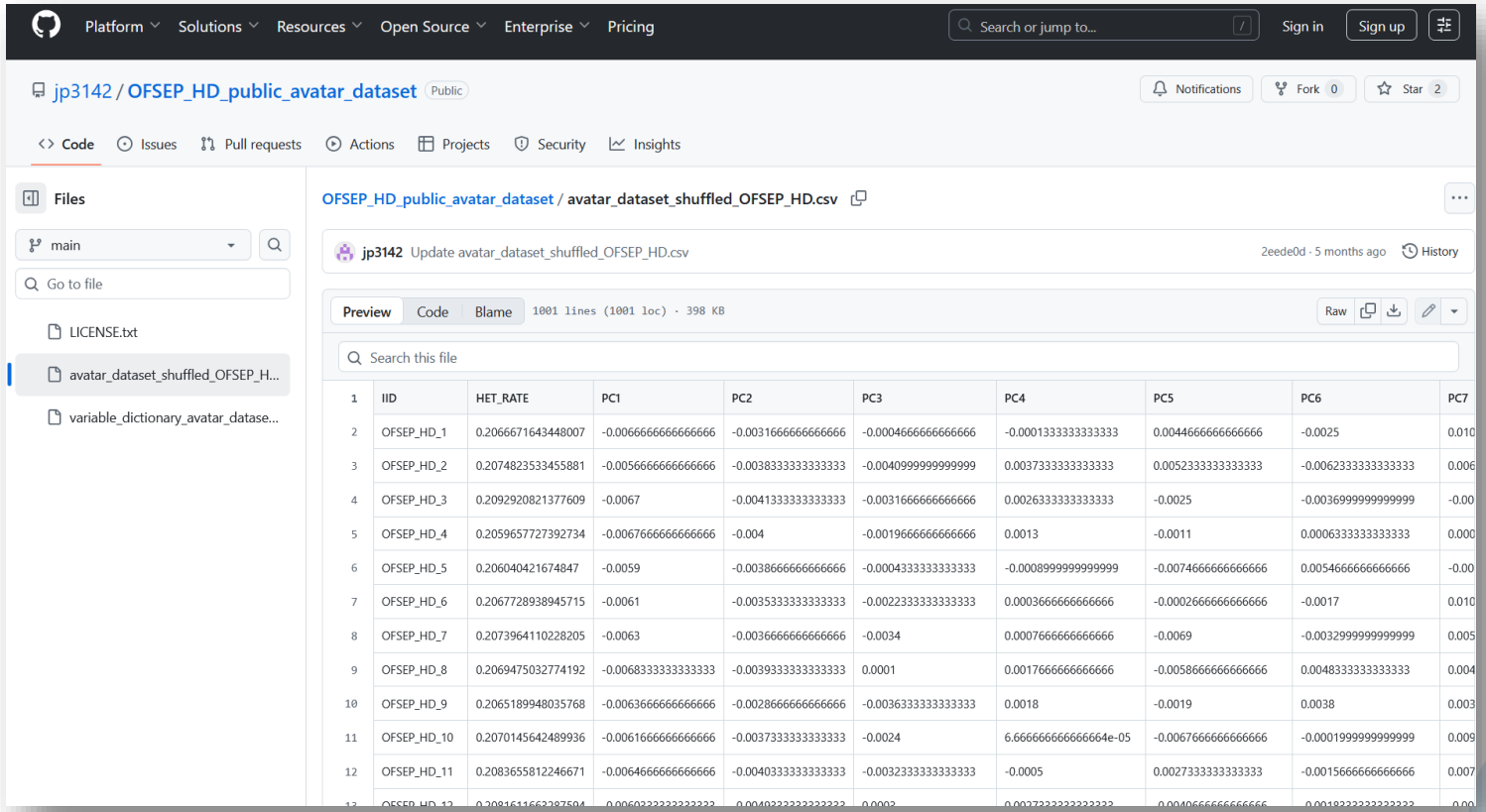


290

291 ***Figure 1:*** *Principal Component Analysis (PCA) plot of the OFSEP-HD population based on the first two principal components*

292 *(PC1 and PC2), overlaid with reference datasets. Each data point represents an individual. Semi-transparent data points*

Innovative Data Strategies: 1-clic access to dataset



	IID	HET_RATE	PC1	PC2	PC3	PC4	PC5	PC6	PC7
2	OFSEP_HD_1	0.2066671643448007	-0.0066666666666666	-0.0031666666666666	-0.0004666666666666	-0.0001333333333333	0.0044666666666666	-0.0025	0.010
3	OFSEP_HD_2	0.2074823533455881	-0.0056666666666666	-0.0038333333333333	-0.0040999999999999	0.0037333333333333	0.0052333333333333	-0.0062333333333333	0.006
4	OFSEP_HD_3	0.2092920821377609	-0.0067	-0.0041333333333333	-0.0031666666666666	0.0026333333333333	-0.0025	-0.0036999999999999	-0.00
5	OFSEP_HD_4	0.2059657727392734	-0.0067666666666666	-0.004	-0.0019666666666666	0.0013	-0.0011	0.0006333333333333	0.000
6	OFSEP_HD_5	0.206040421674847	-0.0059	-0.0038666666666666	-0.0004333333333333	-0.0008999999999999	-0.0074666666666666	0.0054666666666666	-0.00
7	OFSEP_HD_6	0.2067728938945715	-0.0061	-0.0035333333333333	-0.0022333333333333	0.0003666666666666	-0.0002666666666666	-0.0017	0.010
8	OFSEP_HD_7	0.2073964110228205	-0.0063	-0.0036666666666666	-0.0034	0.0007666666666666	-0.0069	-0.0032999999999999	0.005
9	OFSEP_HD_8	0.2069475032774192	-0.0068333333333333	-0.0039333333333333	0.0001	0.0017666666666666	-0.0058666666666666	0.0048333333333333	0.004
10	OFSEP_HD_9	0.2065189948035768	-0.0063666666666666	-0.0028666666666666	-0.0036333333333333	0.0018	-0.0019	0.0038	0.003
11	OFSEP_HD_10	0.2070145642489936	-0.0061666666666666	-0.0037333333333333	-0.0024	6.666666666666664e-05	-0.0067666666666666	-0.0001999999999999	0.009
12	OFSEP_HD_11	0.2083655812246671	-0.0064666666666666	-0.0040333333333333	-0.0032333333333333	-0.0005	0.0027333333333333	-0.0015666666666666	0.007

Uses of Synthetic Data

1. **Anonymization & Transparency** – safe 1-1 exports, auditability
2. **Data Sharing** – internal (students/collaborators), external (open science, sandbox)
3. **Training & Education** – realistic practice, ethical learning (*never first on real patients/data*)
4. **Archiving** – preserve data assets instead of deletion
5. **Augmentation** – calibration, bias correction, sensitivity tests
6. **Decision Support** – safe, audited data for AI/ML tools

Paris et al. preprint: https://www.medrxiv.org/content/10.1101/2025.04.28.25326562v2

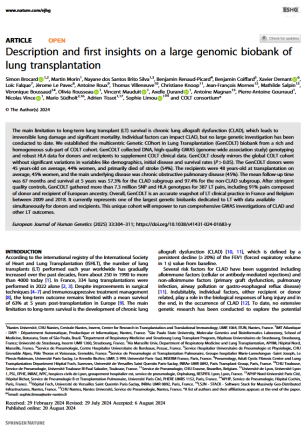
https://github.com/jp3142/OFSEP_HD_public_avatar_dataset/blob/main/avatar_dataset_shuffled_OFSEP_HD.csv

SCIENCE OUVERTE

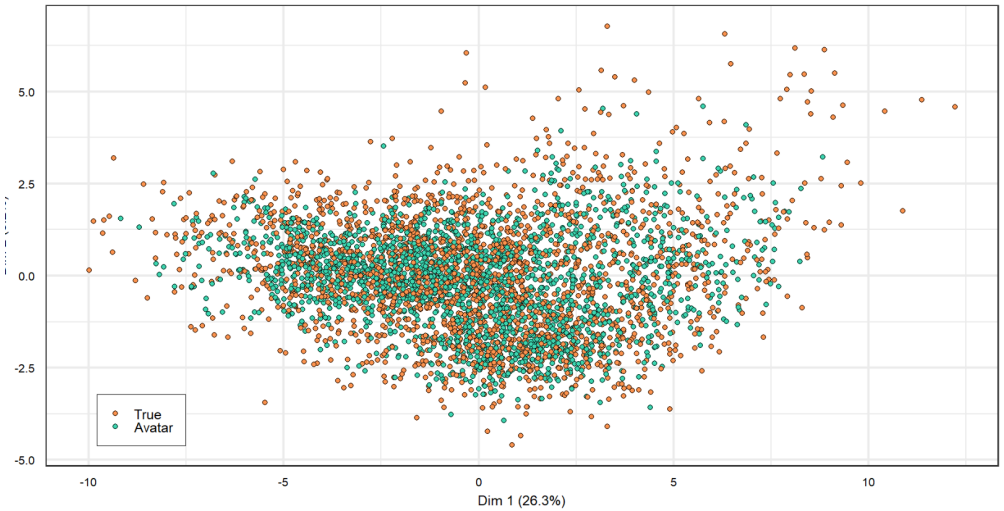
Chercheurs FRBionet & Démarche de publication partielle en opendata.
https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2026-09/MESRE_SSO_2026_FR.pdf.

+14,5 % de citations supplémentaires, en moyenne, pour un article partageant des données

Anonymous Synthetic Data for Transplantation


Description and first insights on a large genomic biobank of lung transplantation


KIT-GENIE, the French genetic biobank of kidney transplantation



center	data	hidden_rat	local_cloaki	distance_to_closest	closest_distances_ratio	column_direct_match_pr	categorical_hidden_rate
Nantes	FAMD	92.1	5	1.611	0.7671	99.75	94.46
Lyon	FAMD	93.32	7	1.895	0.8183	99.66	90.2
Montpellier	FAMD	91.74	7	3.048	0.8948	99.64	94.93
Necker	FAMD	92.03	7	2.39	0.8745	99.81	93.73
Saint_Louis	FAMD	93.52	7	1.633	0.7449	98.76	95.24

DIVAT (cohorte/DPI)

N= 2884 (Nantes)

20 variables & Privacy Metrics

Nantes Université

Conclusion : La technologie au service de la valorisation des Bio collections ?

VALEUR d'USAGE = POTENTIEL BRUT DES DONNEES ou des ECHANTILLONS X TRAVAIL d'ORGANISATION & de TRANSFORMATION

PRIMUS : Données multi-sources, mises en qualité, synthétiques, anonymes, dans une infrastructure distribuée.

Données Synthétiques :

- **Qualité scientifique éprouvée** : Robustes, fiables, testables et validées- « AI-ready »
- **Sécurité juridique et légale** : Certifiée par la CNIL, Conformes au RGPD, respect de la vie privée.
- **Intégrité éthique** : Responsables, transparentes et authentiques.
- **Impact économique et sociétal** : Stimulent l'innovation et favorisent la croissance.
- **Ouverture et accessibilité:** Démocratisent les données pour tous les secteurs.

Nantes Université

PLACE DES DONNÉES ET COLLECTIONS BIOLOGIQUES DANS LES ALGORITHMES DE MÉDECINE DE PRÉCISION

JOURNÉE NATIONALE FRBIONET Le 24 septembre 2026

De la valorisation du réseau OFSEP vers un modèle de médecine de précision dans la sclérose en plaques : Projet RHU PRIMUS

Pr Pierre-Antoine Gourraud

Professeur des Universités & Praticien-Hospitalier

CHU Nantes, PHU 11 : Clinique des données, INSERM, CIC 1413, Nantes Université, INSERM, CR2TI

pierre-antoine.gourraud@univ-nantes.fr

CC BY NC

Nantes Université

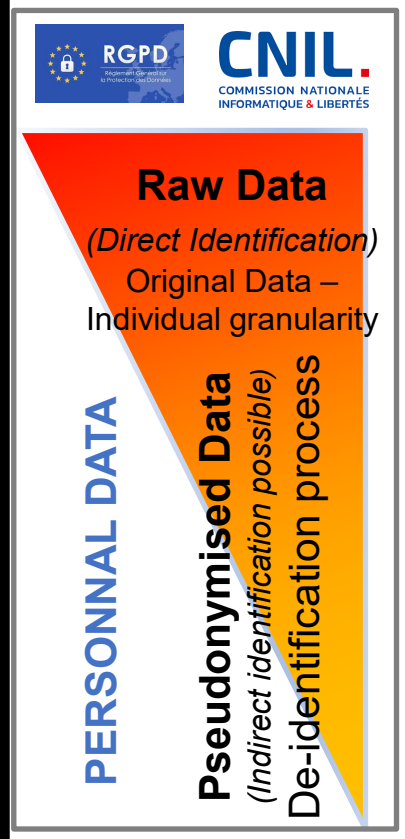
Pôle Santé UFR Médecine

DELPHI DEep Learning Proposal for Healthcare & Innovation AI & Health data

FRANCE 2030 anr Région PAYS de la LOIRE Nantes Métropole

Nantes Université Nantes Université Inserm Institut national de la santé et de la recherche médicale

Graphical conclusion



FIDELITY to real data

N - N - Realistic SD *To learn via data sharing & analytics*

PRIVACY to providers

N=1 Anonymous SD

To respect everyone & unlock data sharing

HYPER-REALITY to individual/phenomena

Hyper-Realistic SD

Exhausting the commensal potential of living phenomea

ANONYMOUS AUGMENTED SYNTETIC DATA

UTILITY to Algorithms

N x X - Augmented SD

To build AI-driven algorithms

Data-Driven Synthetic data generation methods (semi-)Local Modeling

Data-driven & Model-Based Synthetic data generation methods Global models

 # TRUST # Social value of Synthetic Data *To create value for all*

- entre la mise en valeur de la cohorte COLT RHU KTD INNOV
- - et celle de l'OFSEP avant de montrer un exemple de medecine de précision.
- PRIMUS ACCES aux données 3 ingrédients : le travail des donénes comme des échantillons
- La légitimité du contrôle de usage le federé
- Les données comme alpha et oméga
- Pas que la sep -> la genet et COLT et KITgenie