

Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis

The anonymous Synthetic data Era

Session Two C | Future Directions and Opportunities

Introduction and Overview: AI/ML Innovations in Radiation Occupational Health
-2025 March 13th 1:45 pm

NATIONAL ACADEMIES Sciences Engineering Medicine
NUCLEAR AND RADIATION STUDIES BOARD



Gilbert W. Beebe Symposium on AI and ML Applications in Radiation Oncology, Medical Diagnostics, and Occupational and Public Safety

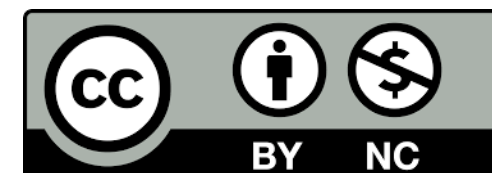
The National Academy of Sciences Building
Fred Kavli Auditorium
2101 Constitution Ave NW
Washington, DC 20418



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



Overview

“Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis : [The anonymous Synthetic data Era](#)”

- **Summary :** Synthetic data generation methods promote collective intelligence and enable sharing codes that apply seamlessly to both original and synthetic data. The use of synthetic data allows unleashing personal data potential to improve future healthcare systems and beyond while ensuring individual privacy. Using a local private model, the Avatar method respects GDPR constraints by enabling data sharing without compromising privacy. Personal data should be restricted to personal use. Anonymous synthetic data generation open new perspectives of data sharing and data augmentation in occupational health.

Disclaimer

- **Potential COI :**

PA Gourraud is the founder of Methodomics (2008) and the co-founder of Big data Santé (2018). He consults for major pharmaceutical companies, and start-ups, all of which are handled through academic pipelines (AstraZeneca, Amgen, Biogen, Boston Scientific, Cemka, Cook, Docaposte/Heva, Edimark, Ellipses, Elsevier, Grunenthal, Janssen, IAGE, Lek, Methodomics, Merck, Mérieux, Novartis, Octopize, Sanofi-Genzyme, Lifen, TuneInsight, Aspire UAE). 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.

Opening

- « ***The United States invents and innovates, China imitates and replicates ... while Europe regulates***»

→ *Let me show you how*

“EU regulation can foster world-wide innovation that replicates individual data”

EU GDPR ... *can foster innovation*



- **The notion of “Personal data” is defined in Art. 4 (1).**
 - Personal data are any information which are related to an identified or identifiable natural person.
- **Article 5 1. on processing of personal data and minimization principle**

1. Personal data shall be:

- (a) ...('lawfulness, fairness and transparency')
- (b) *collected for specified, explicit and legitimate purposes and not further processed (..) statistical purposes shall, in accordance with [Article 89\(1\)](#), not be considered to be incompatible with the initial purposes ('purpose limitation');*
- (c) *adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization');*
- (d) (...) ('Accuracy')
- (e) *kept in a form which permits identification of data subjects for no longer than is necessary for the purposes (...) statistical purposes in accordance with [Article 89\(1\)](#) (...) ('storage limitation');*
- (f) *processed in a manner that ensures appropriate security of the personal data, (...) , using appropriate technical or organizational measures ('integrity and confidentiality').*

- Right of identifiable person : Trust & transparency **GDPR Article 12** and following ...
- **Article 89** GDPR Safeguards and derogations relating to **processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes**

National/EU Regulation GDPR

... *can foster innovation*



https://www.cnil.fr/sites/default/files/atoms/files/referentiel_entrepot.pdf

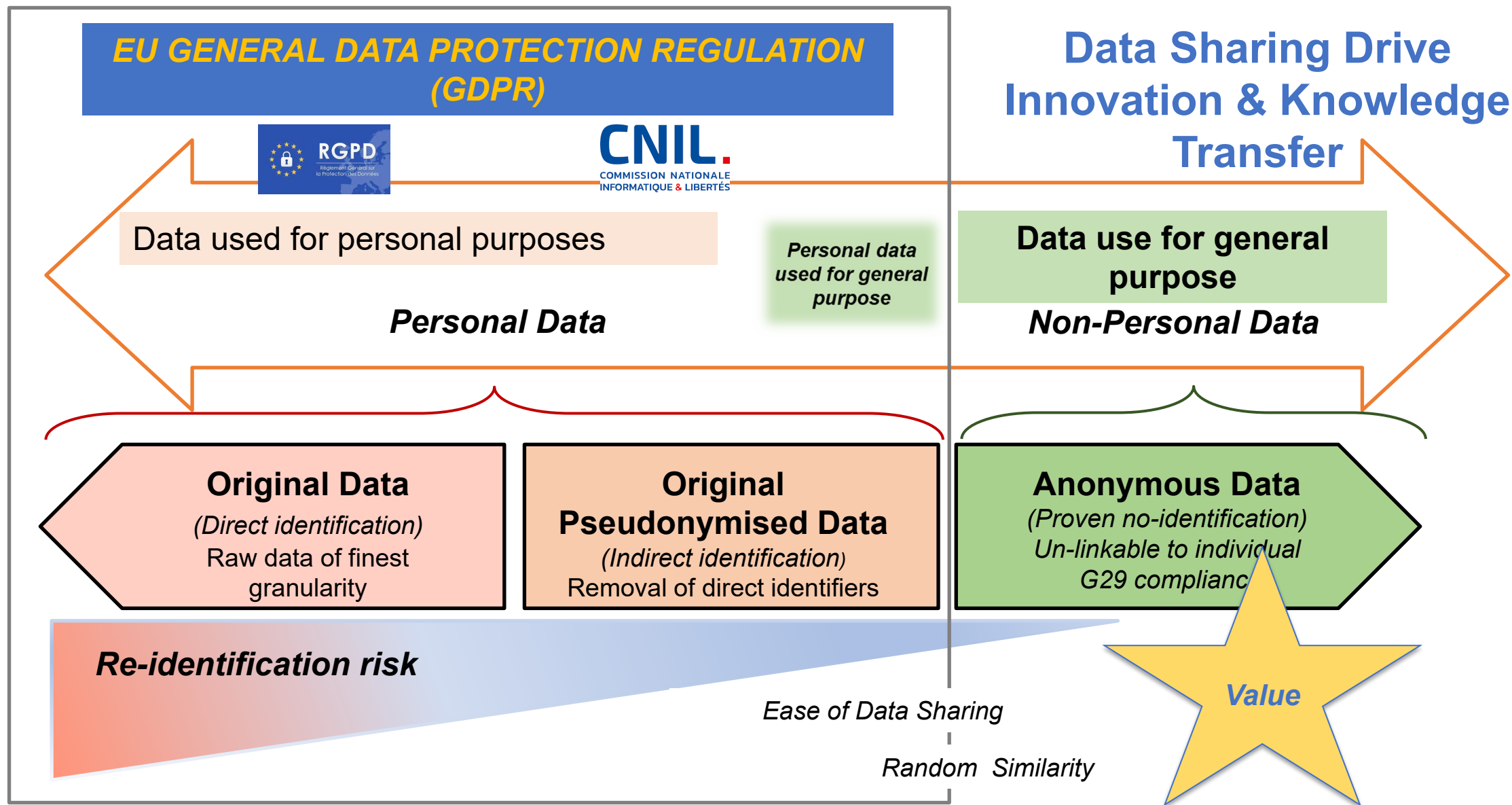
Page 14 : Exporting data from the warehouse and workspaces

SEC-EXP-1 With the exception of data relating to re-identification procedures SEC-REI-1 to SEC-REI-3, **only anonymized datasets may be exported out** of the warehouse or workspace. **The anonymization process must produce a dataset that complies with the three criteria defined by G29 Opinion No. 05/2014** or any subsequent EDPS opinion on anonymization. **This compliance must be documented and demonstrable.** Failing this, if these three criteria cannot be met, a study of the risks of re-identification must be carried out and documented.

SEC-EXP-2 Data exports must be subject to prior **validation by a manager**, in order to endorse the principle, particularly with regard to the SEC-EXP-1 requirement.

SEC-EXP-3 Exports must be monitored automatically or manually by a specialized operator to ensure that they are anonymous. In the case of automatic monitoring, any export identified as non-compliant must be the subject of an alert and quarantine in the warehouse, then manually checked by a specifically trained and empowered manager.

SEC-EXP-4 The systems implemented in the warehouse relating to the production of indicators and the strategic steering of a healthcare facility's activity must only allow anonymous restitutions, including taking into account the filtering and selection functionalities of these restitutions. This restitution process must comply with the three criteria defined by G29 opinion no. 05/2014 or any subsequent EDPS opinion on anonymization. This compliance must be documented. Otherwise, if these three criteria cannot be met, **a study of the risks of re-identification must be carried out and documented.**



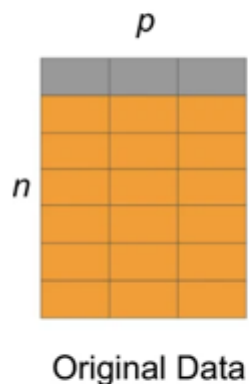
Inspired from E Boucher de Crevecoeur (CNIL) & E Jacqmarcq (Octopize)

Synthetic Data will get popular : 3 *properties*

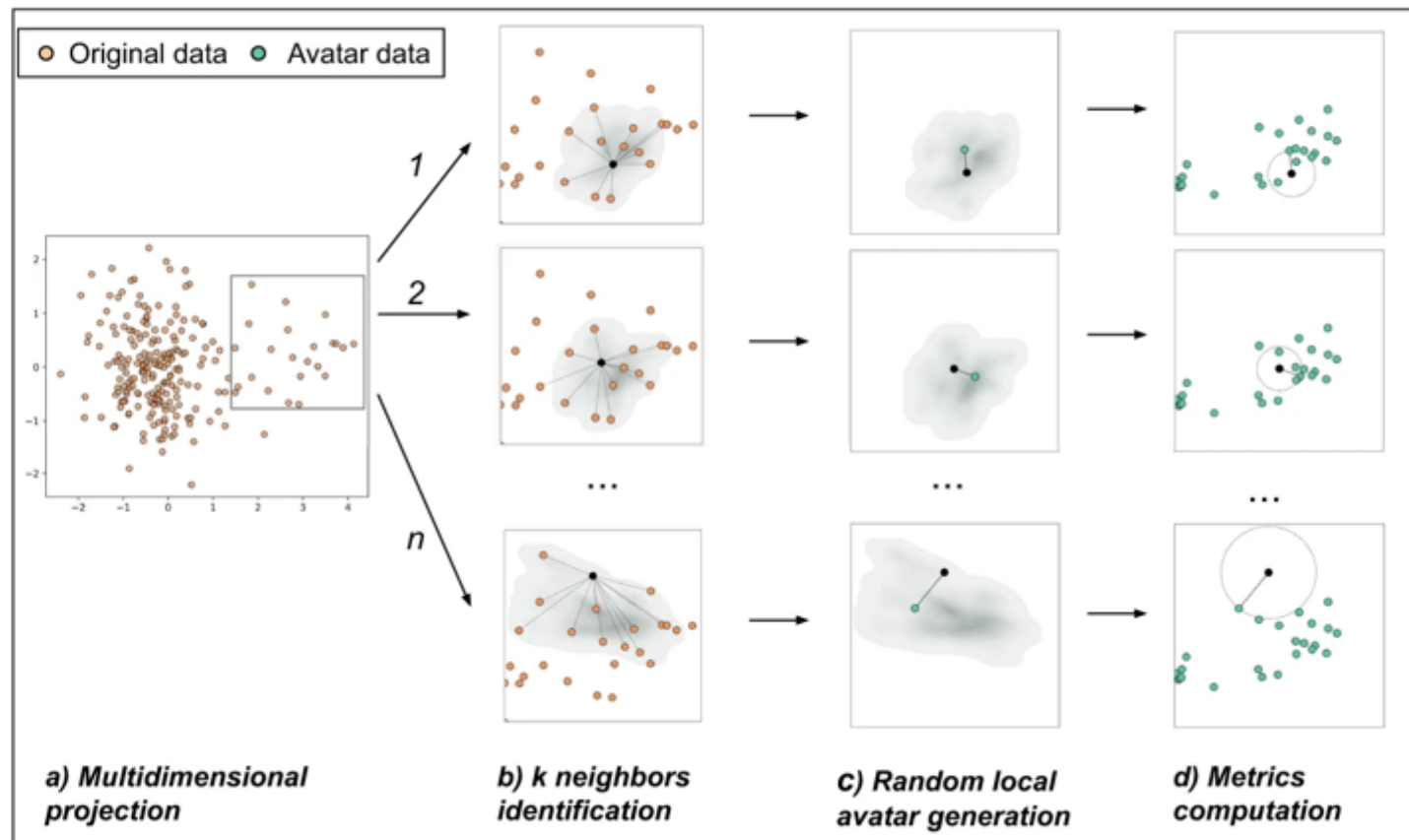
- **(1) Structural similarity**
 - (i.e., the same granularity) : same statistical unit - generic vehicle of information
 - *The synthetic dataset contains the same number of observations, the same number of variables and the same variable types;*
- **(2) Information relevance**
 - A data analyst will obtain results from the synthetic dataset that are comparable to the original data. It make sense.
- **(3) Subjective assessment**
 - Neither experts, nor trained algorithms can distinguish synthetic data from original data.
- **Synthetic Anonymous data** (ARTICLE 29 DATA PROTECTION WORKING PARTY - **G29 Opinion No. 05/2014 EDPB**)
 - Document : (i) is it still possible to single out an individual ?*
 - Document : (ii) is it still possible to link records relating to an individual ?*
 - Document : (iii) can information be inferred concerning an individual ?*
 - Yes? No? Maybe? – how much ?*

HOW ? Privacy-by-design Anonymous synthetic data generation

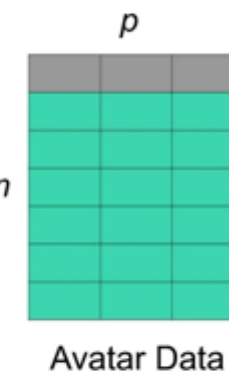
1- Input



2- Avatar method



3- Output



Guillaudeau et al 2023
Nature Digital
Medicine

npj | Digital Medicine



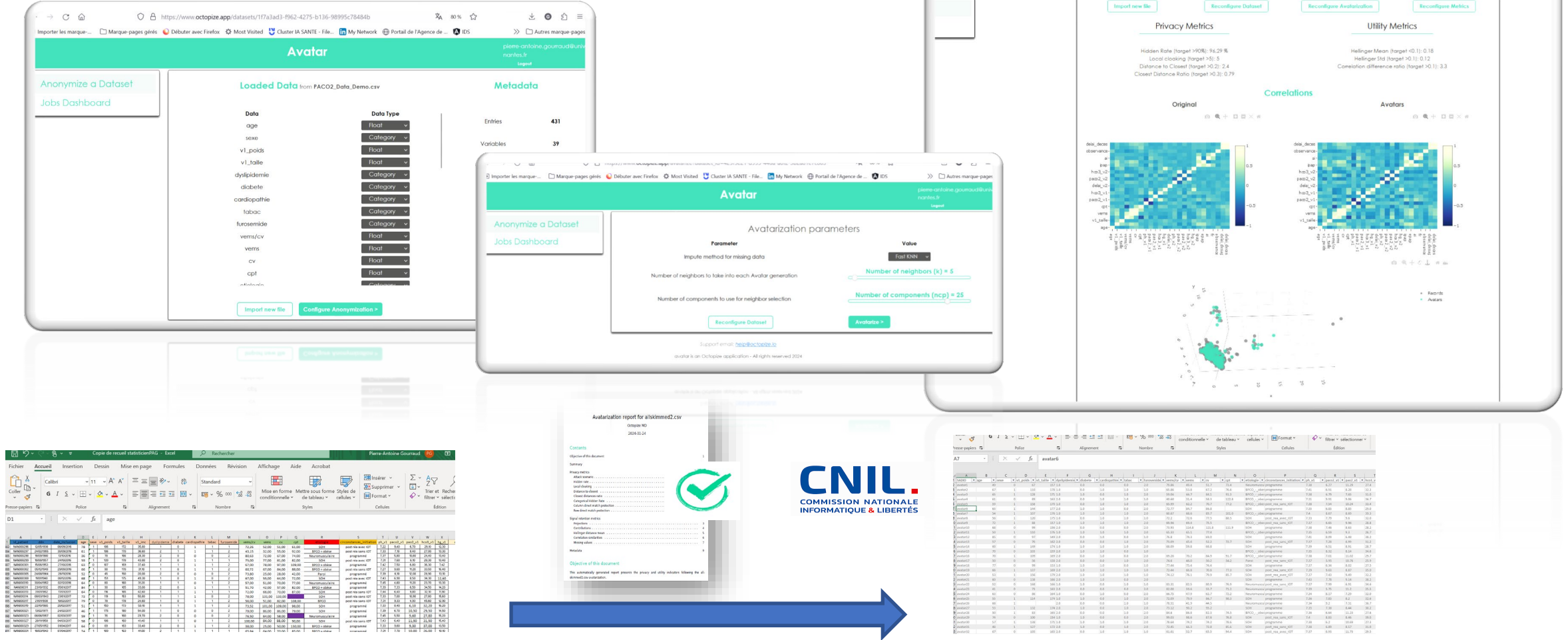
<https://www.nature.com/articles/s41746-023-00771-5>

Use Case : Data Clinic & Open Sciences

- **Context :** *Pr FX B and DR S J ask the Data Clinic for help in responding to reviewers, the last hurdle to publication entitled : “Does PaCO₂ correction have an impact on survival of patients with chronic respiratory failure and long-term non-invasive ventilation?” (Thomas et al.2024)*
- **Comment 1 – rev 1 :** Study with lack of statistical power, clinical heterogeneity, possibility of sharing data (for meta-analysis)?
 - *Pseudonymised Data – Data Re-use Control ?*
- **Solution :** Anonymous Synthetic Data
 - National Reference guide for BDW
 - Simulate Data out of GDPR
- **Other example :** JNNP open publication code & data (MIT License)
 - <https://jnnp.bmj.com/content/92/2/122>
 - <https://github.com/ICAN-aneurysms/RIA-predict/tree/master/notebooks>



Live Demo



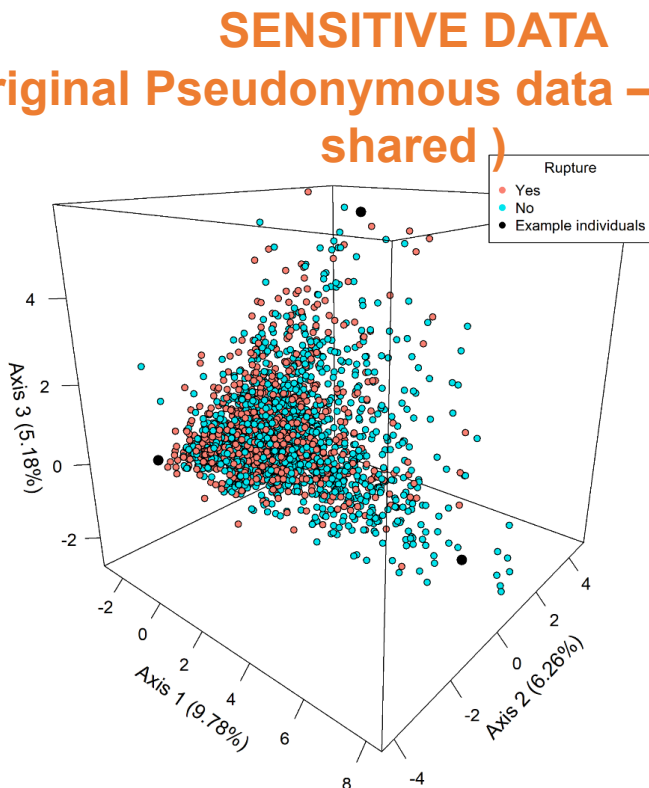
<https://doi.org/10.1016/j.heliyon.2024.e26437>

<https://ars.els-cdn.com/content/image/1-s2.0-S240584402402468X-mmc1.xlsx>

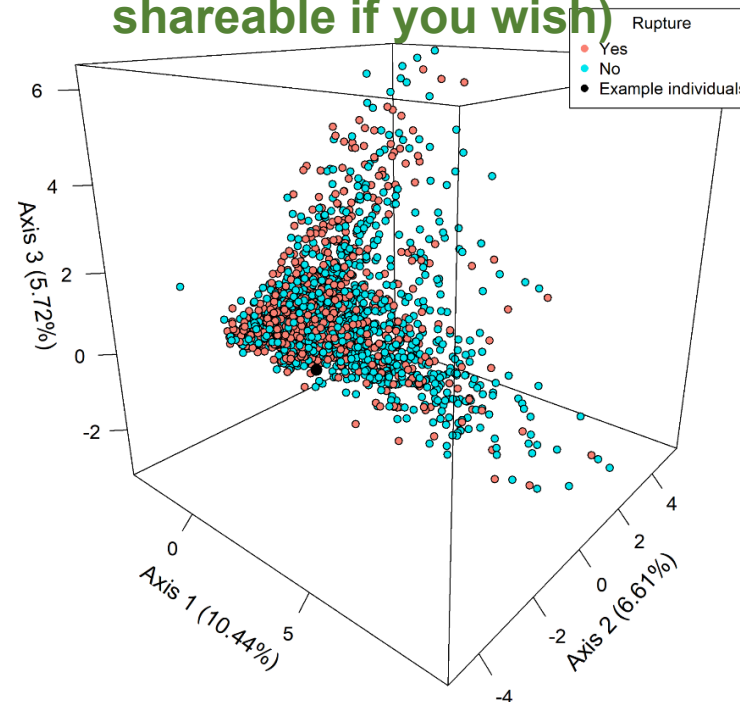
Comparison of Factorial Analysis of Mixed Data

AVATARS

(synthetic version data the original dataset - shareable if you wish)



<https://jnnp.bmj.com/content/92/2/122>
<https://github.com/ICAN-aneurysms/RIA-predict/tree/master/notebooks>



CNIL.

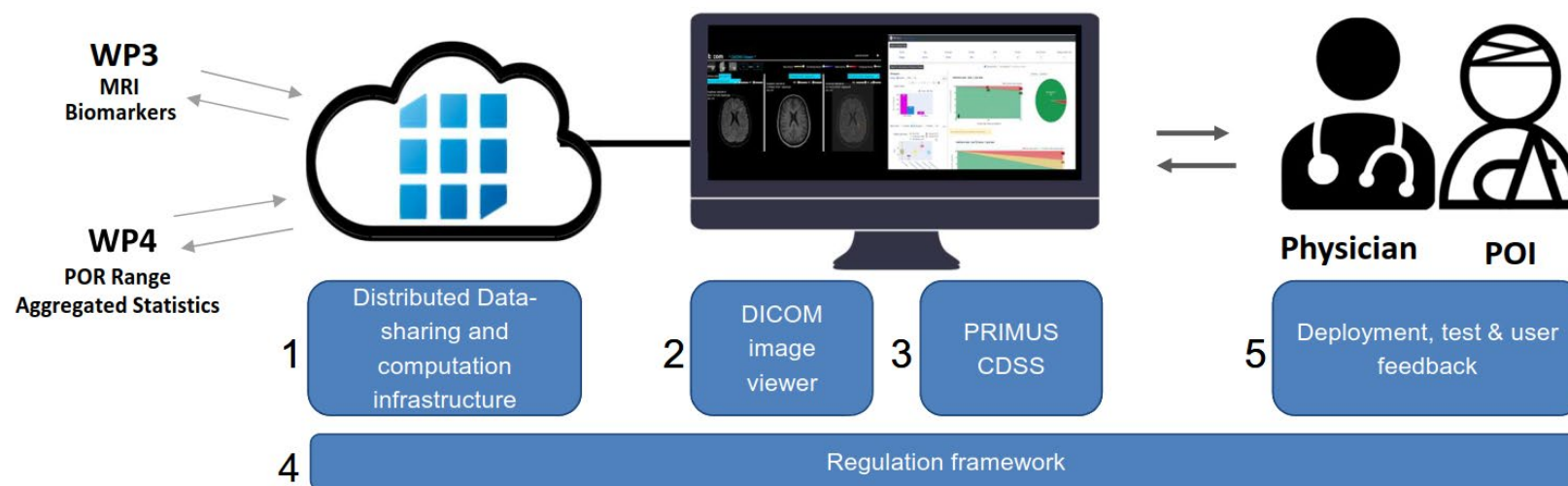
SENSITIVE DATA (Original Pseudonymous data)	
Risk of re-identification	+++
Preparation/curation	+
Informative Value	++++

AVATARS (synthetic version data the original dataset)	
Risk of re-identification	0
Preparation/curation	+
Informative Value	++++

Primus Project in Multiple Sclerosis – overview

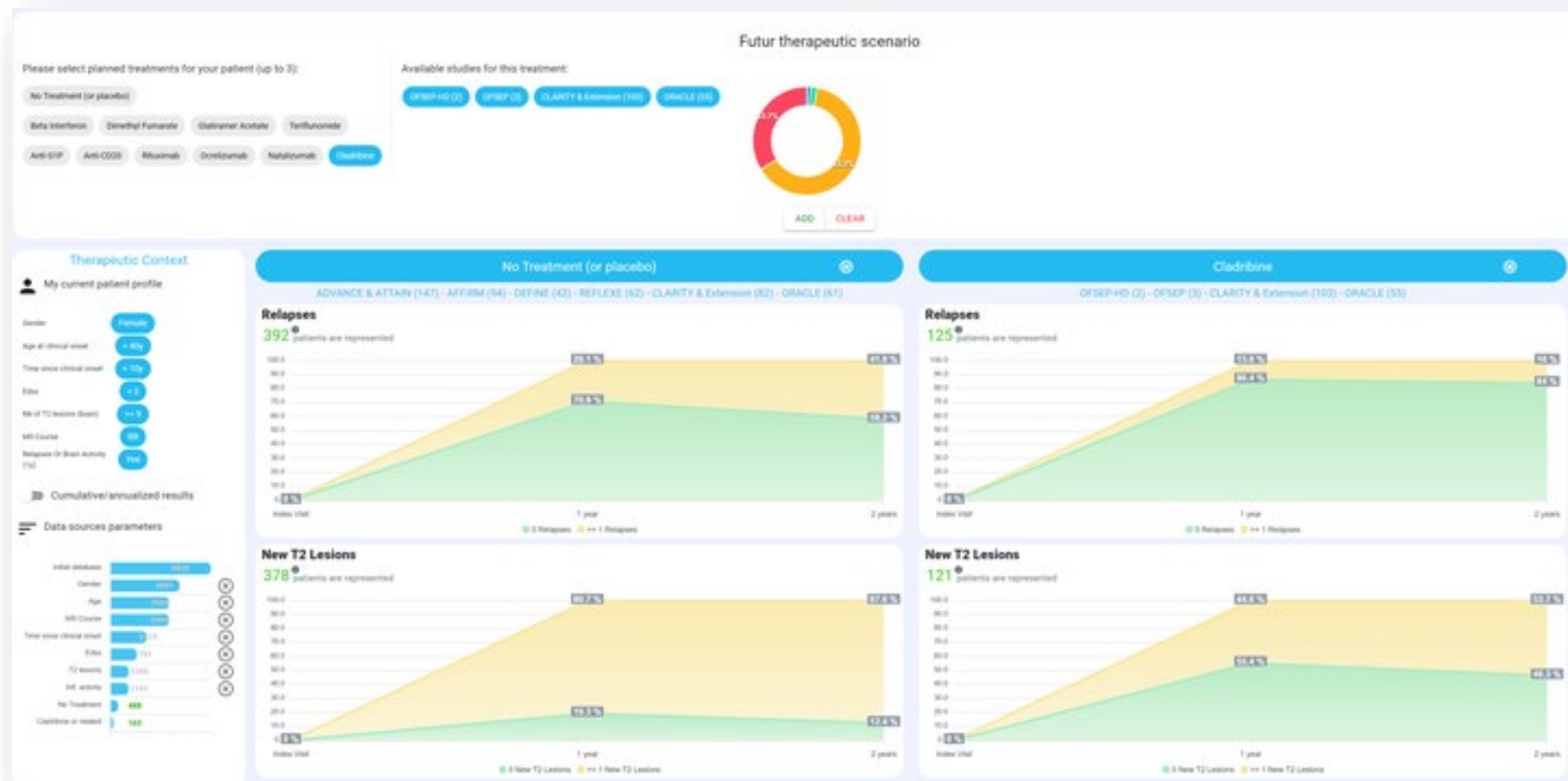
Clinical Decision Support Software (**CDSS**) that enable neurologists and patients (Patient Of Interest - **POI**), during a consultation to visualize (**viewer**) the projection of the POI's disease progression in relation to Reference data using an array of collaborative algorithms (**AI**) that will select reference patients (**POR**) from a reference database (**BD Ker PRIMUS**) built from industry clinical trials and real-life data in compliance with regulations (**RGPD**) and the Medical Device Regulation (**MDR**). Tested in a clinical trial in preparation for **CE marking**.

"The right treatment for the right person at the right time and prevent disability".



// What ? *A clinical decision support System (CDSS) based on Data & algorithms*

Data-driven comparison of therapeutic scenarios (WP5 prototype)



// What ? A clinical decision support System (CDSS) based on Synthetic Data & algorithms

Selection criteria

At the current visit:

- ☒ **MS course**
- RIS (0 ; 0) ☐
 - RR-MS (196 ; 79) ☒
 - SP-MS (0 ; 0) ☐
 - PP-MS (0 ; 0) ☐

- ☒ **Age at clinical onset**
- < 26y (0 ; 0) ☐
 - 26 - 40y (196 ; 79) ☒
 - ≥ 40y (0 ; 0) ☐

- ☒ **Disease duration**
- < 2y (190 ; 79) ☒
 - 2 - 10y (0 ; 0) ☐
 - ≥ 10y (0 ; 0) ☐

- ☒ **EDSS**
- 0 - 2.5 (196 ; 79) ☒
 - 3 - 5.5 (0 ; 0) ☐
 - 6 - 10 (0 ; 0) ☐

- ☐ **Sex**
- Female (148 ; 48) ☒
 - Male (48 ; 31) ☐

- ☒ **Pregnancy status**
- Not pregnant (196 ; 79) ☒
 - Pregnant/Planned (0 ; 0) ☐

- ☐ **Number of brain T2 lesions**
- < 9 (41 ; 17) ☐
 - ≥ 9 (141 ; 46) ☒

- ☐ **Number of spinal T2 lesions**
- 0 (24 ; 21) ☒
 - ≥ 1 (33 ; 41) ☐

During the previous year:

- ☐ **1) Number of relapses**
- 0 (34 ; 17) ☒
 - 1 (136 ; 52) ☐
 - ≥ 2 (29 ; 10) ☐

- ☒ **2) New T2 or GdE brain lesions**
- 0 (0 ; 0) ☐
 - ≥ 1 (196 ; 79) ☒

- ☐ **3) Disability evolution**
- Stable or improving (113 ; 59) ☒
 - Sustained worsening (32 ; 8) ☐

- ☐ **(1 or 2) Any relapse or MRI activity**
- Stable (0 ; 0) ☐
 - Active (196 ; 79) ☒

- ☐ **(1 or 2 or 3) Evidence of disease activity**
- NEDA-3 (0 ; 0) ☐
 - EDA-3 (0 ; 0) ☒

Past treatment

No Treatment

- Filter by class (196 ; 79) ☐
- Filter by efficacy level (196 ; 79) ☒
- Treated or not (196 ; 79) ☐
- Any scenario but maintenance (0 ; 79) ☐

Planned treatment:

☒ Planned treatment

- Option 1

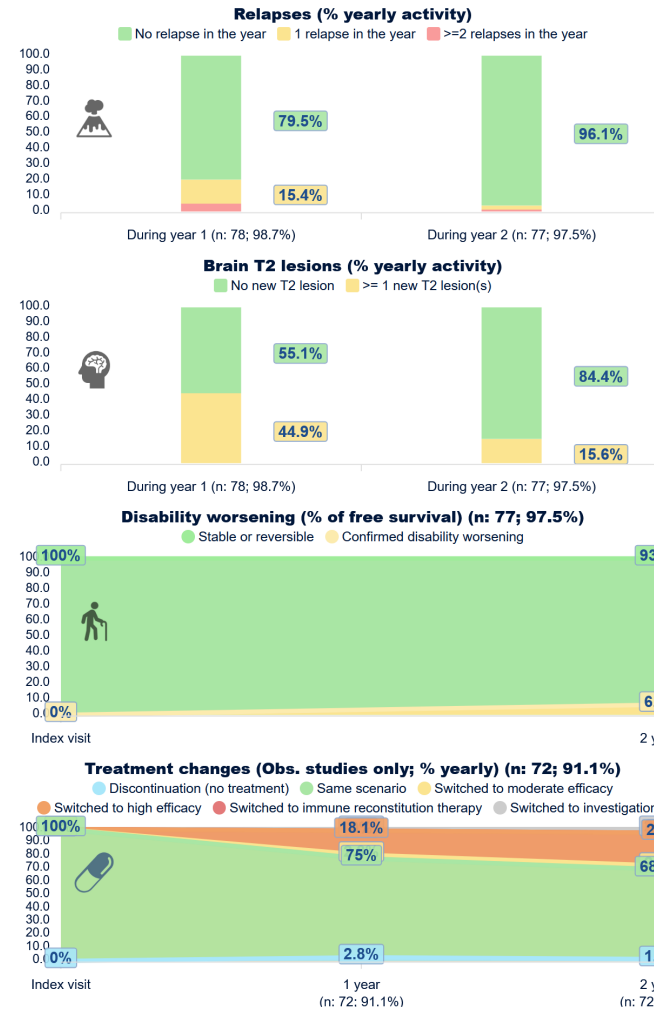
No Treatment

- Option 2

Dimethyl Fumarate or related

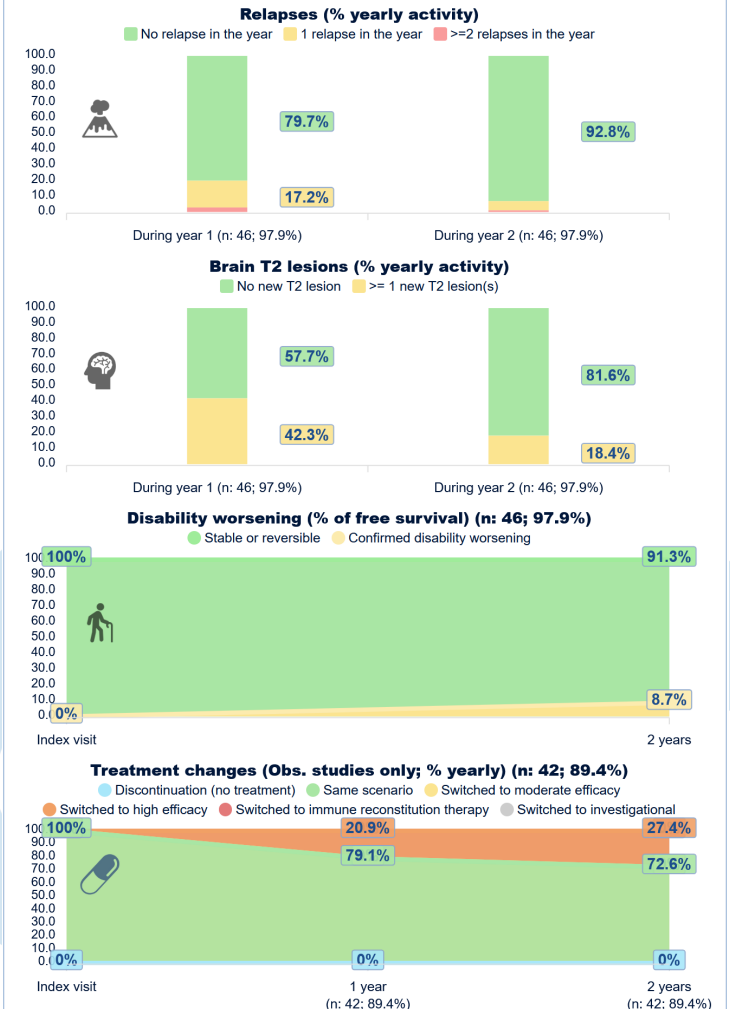
Real data

Dimethyl Fumarate or related (79 patients)



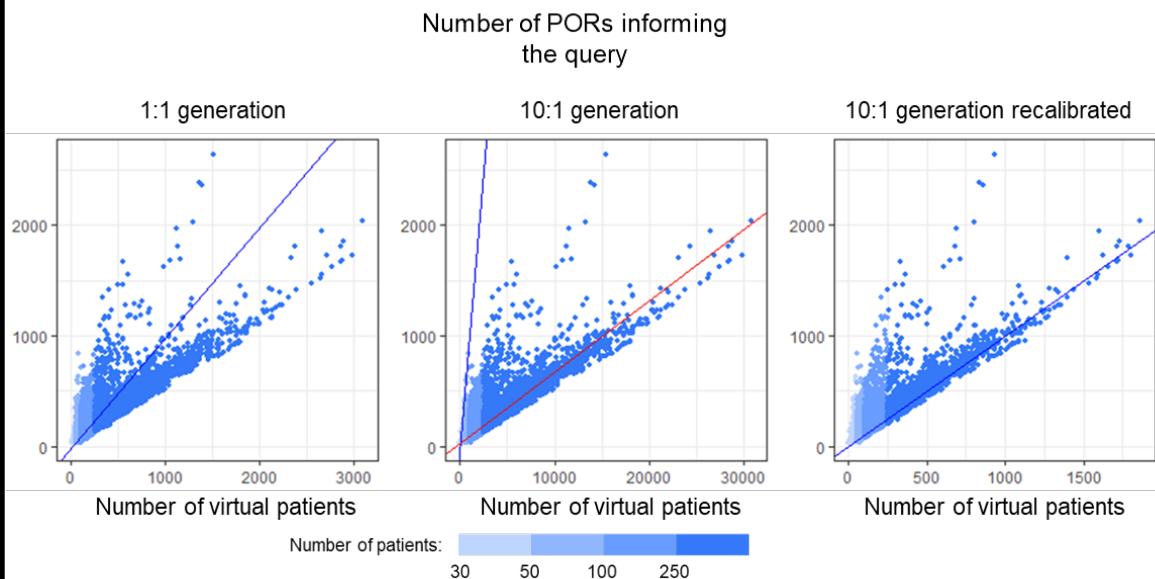
Virtual patients

Dimethyl Fumarate or related (47 patients)



Results: BAKP of the PRIMUS database

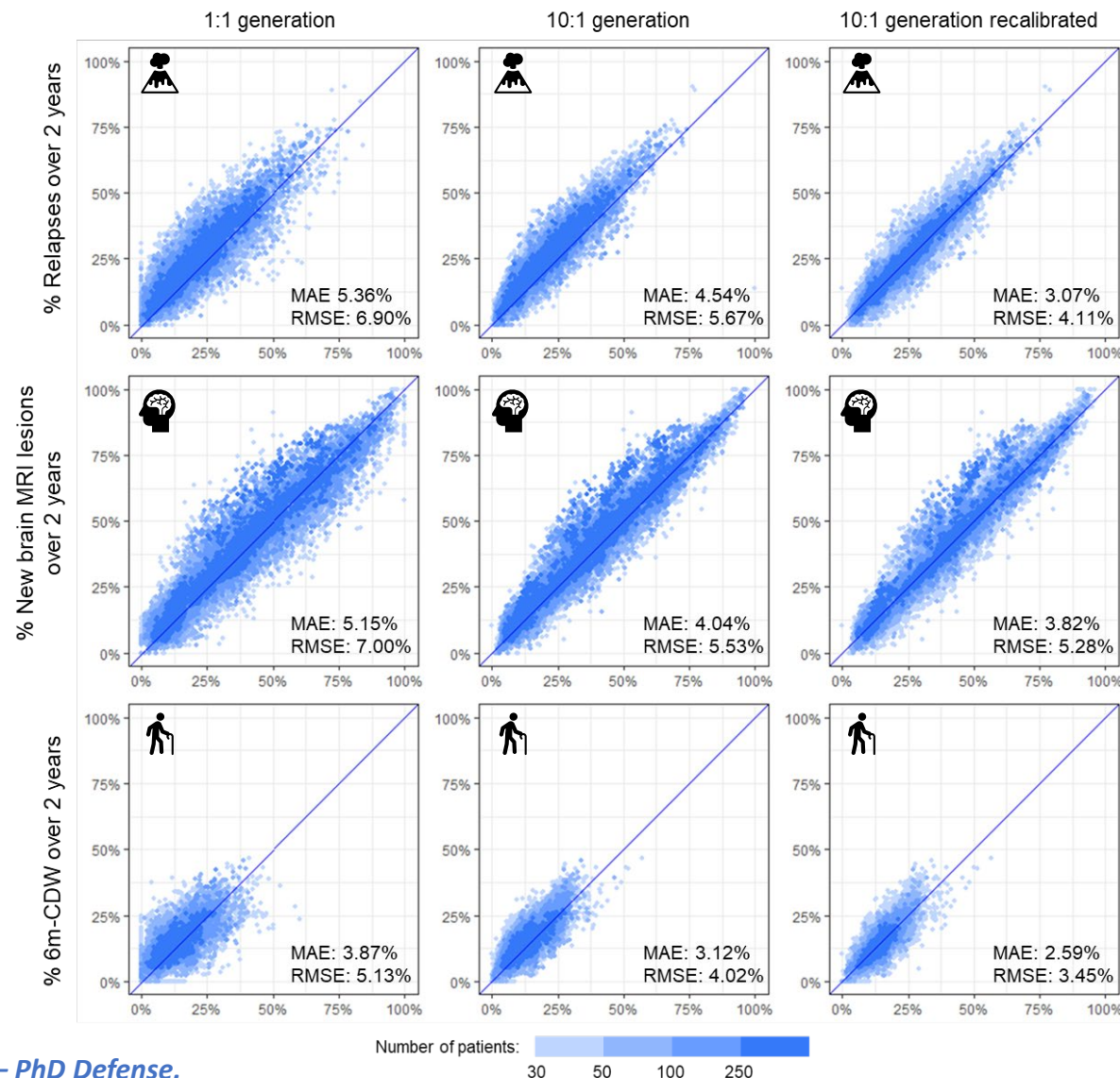
UTILITY (5/5)



6m-CDW: 6-month-confirmed disability worsening

MAE: Mean absolute error

RMSE: Root mean square error



Stanislas DEMUTH – PhD Defense,
Strasbourg, December 18th, 2024

Conclusion : Non Mutually exclusive Uses cases of anonymous synthetic (health) Data

1. **Anonymization perspective** (*Export from Biomedical Warehouse*)
 - (N times; 1-1)
2. **Data sharing** ($N' < ?N$; 1-1) Multi-usages contextual generation
 - Internal : Data sharing
 - External : Innovation and hands' on training (Students, Collaborators, Consortium)
 - “ **Sandbox**” - With private partners – Monetization
3. **Teaching /Training know-how transfer**
4. **Software testing**
 - Ethical issue : 'No first-time procedures on patients.
5. **Archiving** “data capital” instead of deleting.
6. **Data augmentation** ($N' > ?N$; 1-N)
 - Calibration
 - Regularization
7. **Decision Support System** – with safe and enhanced Data

Conclusion :

“This message comes to you today from an old country, France, from an old continent like mine, Europe, that has known wars, occupation and barbarity. A country that does not forget and knows everything it owes to the freedom-fighters who came from America and elsewhere. And yet has never ceased to stand upright in the face of history and before mankind. Faithful to its values, it wishes resolutely to act with all the members of the international community. It believes in our ability to build together a better world. “

D De Villepin 2003 UNC NY, 02/14, 2003

- **“EU regulation can foster world-wide innovation that replicates individual data”**
- Multi-modal anonymous synthetic data : Images, Signals (ECG, GPS, etc) , Genetics Texts
- Privacy G29 and anonymization becomes central in AI regulation
 - https://www.edpb.europa.eu/news/news/2024/edpb-opinion-ai-models-gdpr-principles-support-responsible-ai_en
- **No reason to risk re-identification in biomedical data analysis**

The Data Clinic Team – January 2024



Pierre-Antoine GOURRAUD
Responsable
Pôle de Santé Publique
Sep. 2015



Matthieu WARGNY
Responsable adjoint
Médecin de Santé publique
Nov. 2018



Thomas GORONFLOT
Epidémiologiste
Déc. 2017



Sandrine COUDOL
Epidémiologiste
Fév. 2018



Matilde KARAKACHOFF
Epidémiologiste
Déc. 2018



Soline BOBET
Epidémiologiste
Fév. 2023



Chloé DOUAREC
Ingénieur data
Nov. 2023



Adrien BAZOGE
Ingénieur TAL
Nov. 2023



Emilie VAREY
Chef de projets / Dpt DPI
Nov. 2021



Delphine TOUBLANT
Chef de projet SI
Déc. 2019



Laëtitia AUBERT
Master MEDS
Oct. 2022

Pacôme CONSTANT DIT BEAUFILS

Doctorant / Neurologue
Nov. 2021



Etudiants M2 RC ou BioInfo

Internes de Spécialité

Acknowledgments



Team 5 « NEMO » NEuroinflammation, Mechanisms and therapeutic Options



Team 3 “iTHINK” integrative Transplantation, HLA, Immunology and geNomics of Kidney injury



Ce travail bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du 3e PIA, intégré à France 2030 portant la référence ANR-21-RHUS-0014

Nous remercions nos partenaires, notamment l'OFSEP, Merck Santé et Biogen qui fournissent les données utilisées.



Prof. Gilles Edan

PRIMUS consortium

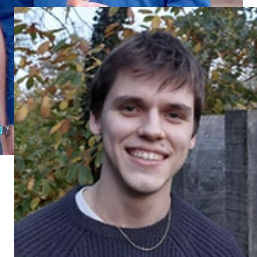
Prof. G. Edan,
Prof. P-A. Gourraud,
Prof. D. Laplaud,
Prof. L. Michel,
Prof. S. Vukusic
Dr. A. Kerbrat
Dr. E. Le Page
Dr. C. Dumas
Dr. R. Casey
N. Blanc
J. Martin-Gauthier
G. Jarre
S. Doyle
A. Auffret
F. Dubois
M. Payet

Prof. P-A. Gourraud,
Dr. S. Limou,
Dr. N. Vince,
Dr. M. Morin,
J. Martin-Gauthier,
J. Paris,
I. Faddeenkov
C. Ed-Driouch,
O. Rousseau,
J. Paris,
I. Faddenkov,
A. Durand,
V. Mauduit,
V. Douillard,
I. Charles,
L. Boussamet,
S. Sayadi,
Nayane,
S. Demuth



Prof. Pierre-Antoine Gourraud,
INSERM U1064, Nantes,
"Translational Immunogenomics of
Transplantation and Autoimmunity"

Prof. D Laplaud,
L. Berthelot,
A. Nicot,
F. Cornelis,
E. Dugast,
A. Garcia,
A. Serova-Erard
J. Louët



Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis

The anonymous Synthetic data Era

Session Two C | Future Directions and Opportunities

Introduction and Overview: AI/ML Innovations in Radiation Occupational Health
-2025 March 13th 1:45 pm

NATIONAL ACADEMIES Sciences Engineering Medicine
NUCLEAR AND RADIATION STUDIES BOARD



Gilbert W. Beebe Symposium on AI and ML Applications in Radiation Oncology, Medical Diagnostics, and Occupational and Public Safety

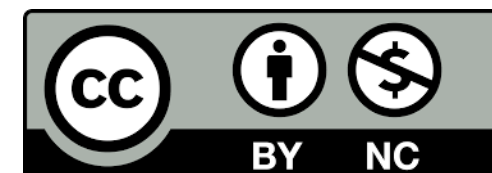
The National Academy of Sciences Building
Fred Kavli Auditorium
2101 Constitution Ave NW
Washington, DC 20418



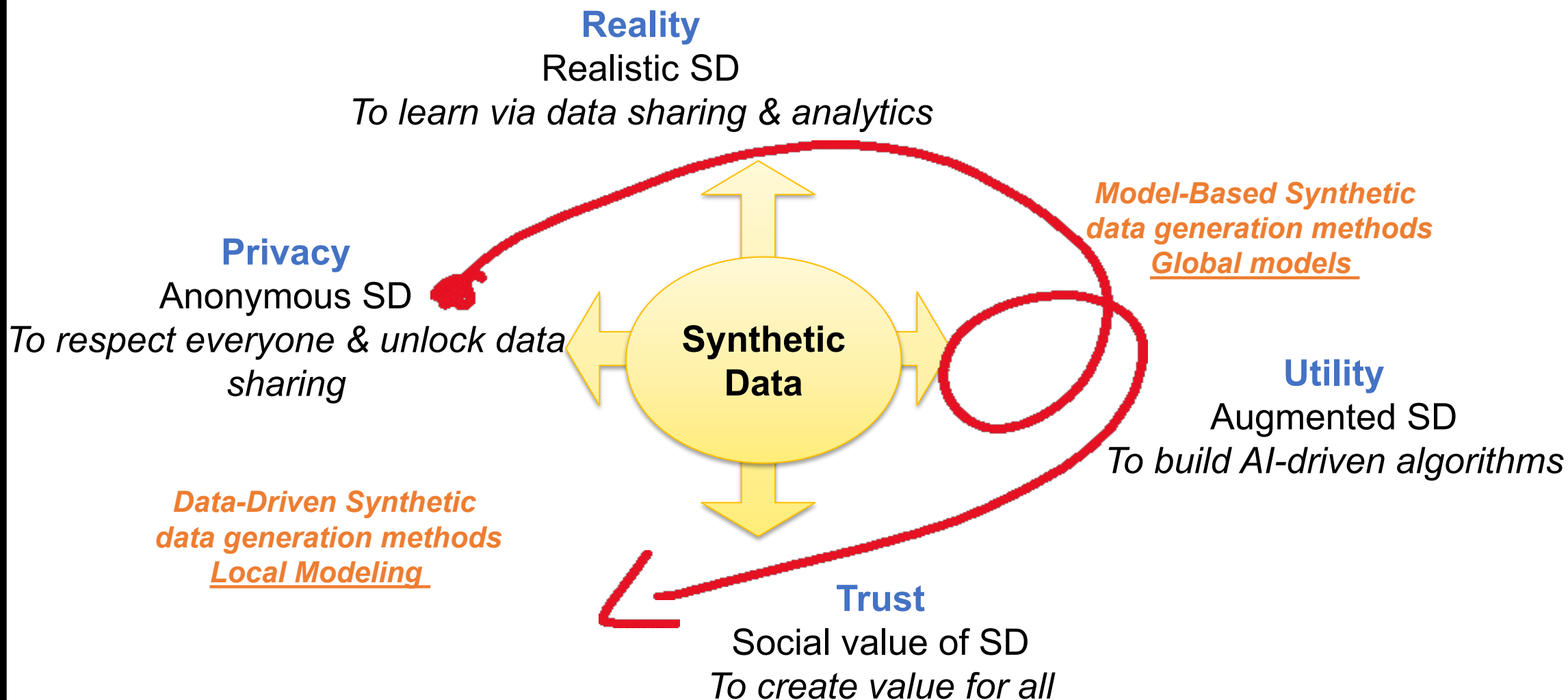
Pierre-Antoine Gourraud

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



Conclusion : A graphical summary of my personal journey in synthetic data world



This work was supported in part by the French National Research Agency (Agence nationale de la recherche – ANR) as its 3rd PIA, integrated to France 2030 plan under reference: ANR-21-RHUS-0014

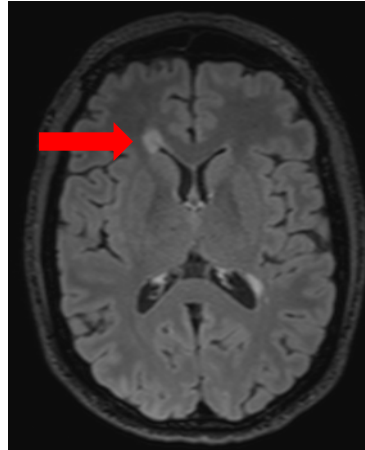


// Why PRIMUS ?

To help person with Multiple Sclerosis (MS)

MS Cohorts

- 112 000 people with MS in France
- 1.8 million worldwide
- MS onset = 30 yrs old



15 drugs

= Data

$$edss_{agg}(x) = \sum_k w_k (floor(edss_k(x))) edss_k(x) \\ \times \left(1 / \sum_k w_k (floor(edss_k(x))) \right)$$



= Algorithms & Computing

« To ensure the right treatment for the right patient at the right time in order to prevent disability »



PRIMUS - Information

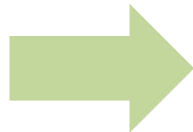
- **PRIMUS** = PRojection In Multiple Sclerosis
- Selected in December 2021
- *RHU = Recherche Hospitalo-Universitaire*
= Including industrial partners & with a business purpose
- Holder = CHU de Rennes
- Coordinator = Pr Gilles Edan
- Scientific Director = Pr PA Gourraud
- Project manager = Nathalie Blanc
- Global cost: 26 439 583 €
- ANR Grant: 8 271 646 € (1/3)
- Duration: 5 years
- Beginning: May 1th 2022
- Ending: April 2028 (Extension ?)





Who ? PRIMUS partnership





// How ? *It is all about data : Treatments Classes (WP2)*

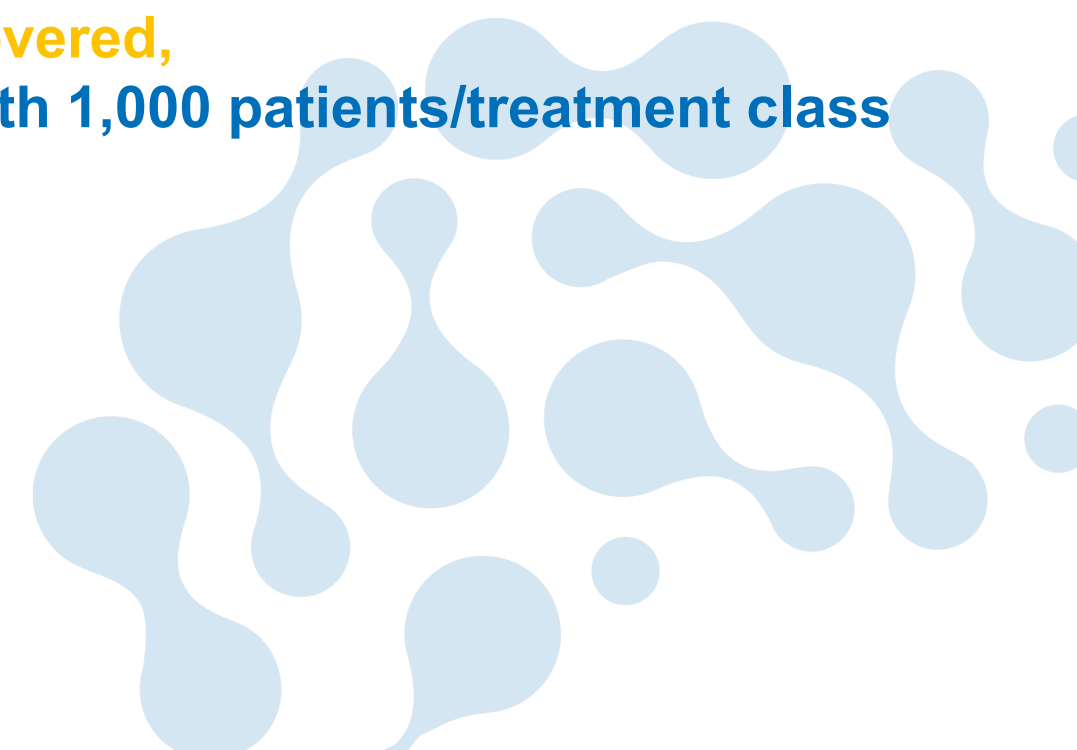
Molecules	Industrials	Commercial	
Beta interferon	Biogen	Avonex, Plégridy	✓
	Merck	Rebif	✓
Dimethylfumarate	Biogen	Tecfidera	✓
Natalizumab	Biogen	Tysabri	✓
Cladribine	Merck	Mavenclad	✓
Glatiramer Acetate	Teva	Copaxone	OFSEP cohort only
Teriflunomide	Sanofi	Aubagio	OFSEP cohort only
Anti CD20	Roche	Ocrevus	OFSEP cohort only
	Novartis	Kesimpta	OFSEP cohort only
	Neuraxpharm	BriumVI	OFSEP cohort only
Anti sphingosine 1 phosphate receptor (S1P)	Novartis (Fingolimod)	Gilenya	OFSEP cohort only
	Janssen /Juvisé Pharmaceuticals (Ponesimod)		OFSEP cohort only

// How ? *It is all about data* : *Beyond RCTs (WP2)*

Classes of treatments	BKP	+ Part of Mother cohort
Beta Interferon	1 533	1 805
Dimethylfumarate	962	1 715
Natalizumab	1 712	3 179
Cladribine	868	900
Glatiramer Acetate	517	875
Teriflunomide	269	926
Anti CD20	490	1 658
Anti sphingosine 1 phosphate receptor (S1P)	386	1 559
	6 737	12 617

= Patients with active disease

All major treatment classes are covered,
with 1,000 patients/treatment class





Howa ? It is all about Data

Randomized Clinical Phase 3 Trials

MERCK

 **Biogen**

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)

Advance

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

Define

- No treatment (n = 408)
- Dimethyl Fumarate (n = 410)

Confirm

- No treatment (n = 363)
- Dimethyl Fumarate (n = 359)
- Glatiramer Acetate (n = 350)

Affirm

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

Sentinel

- Beta Interferon (n = 595)
- Beta Interferon + Natalizumab (excluded)

Ascend

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

Observational studies



HD Cohort

- All treatments (n = 2 842)

Part of Mother Cohort

- All treatments (n = 15 962)

Ker Primus Data Warehouse (BKP)

Data ingestion, harmonization, integration and formatting

N = 25 737 patients of reference



Ker Primus Synthetic Data Warehouse (BAKP)

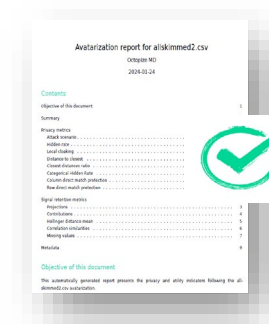
Data ingestion, harmonization, integration and formatting of anonymous synthetic data

N = 25 737 patients of reference

Results: The Privacy-Utility trade-off of Synthetic data → anonymous synthetic Data

PRIVACY (1/1)

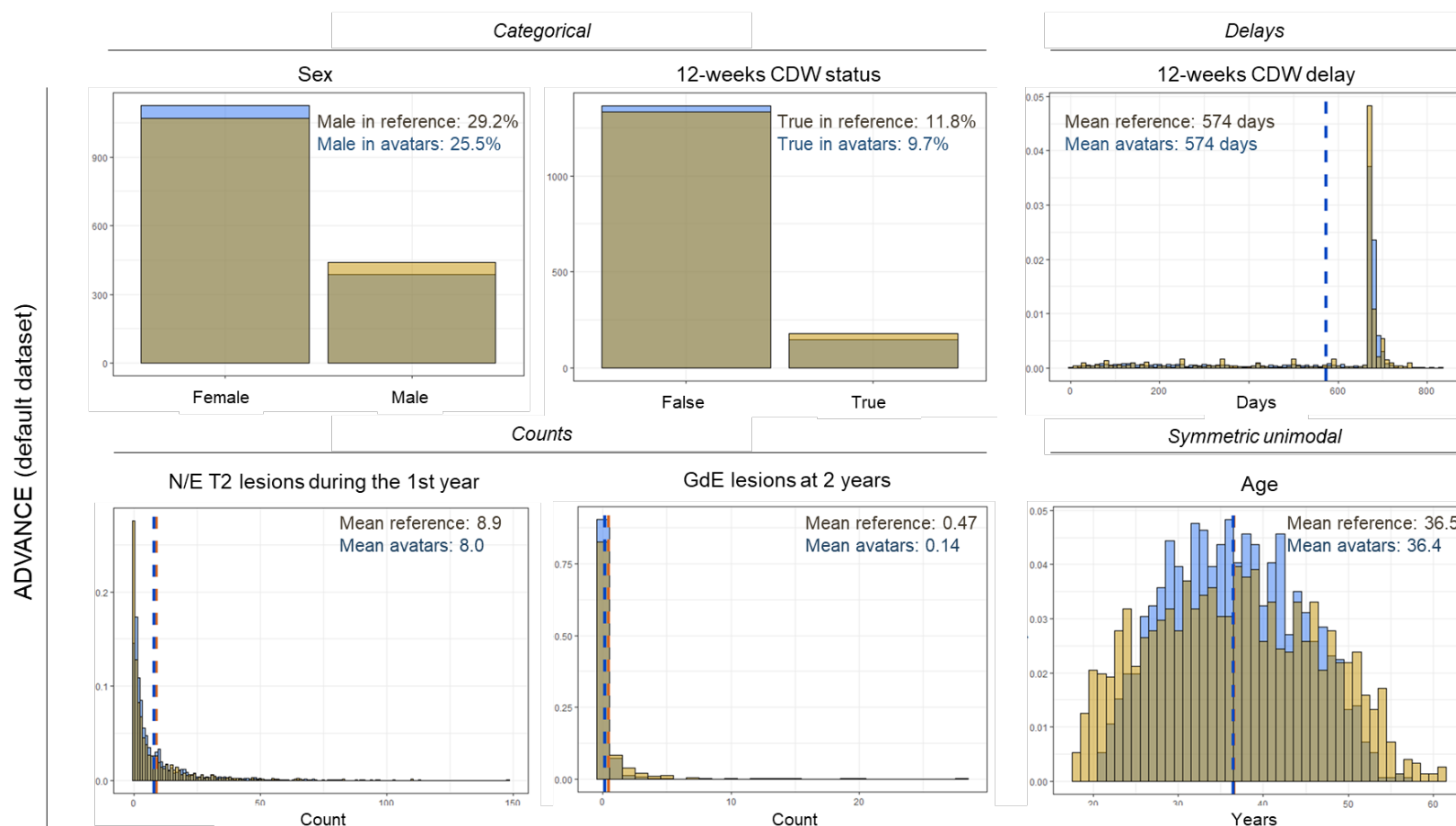
Anonymization G29 criteria	Privacy Metrics	Recommended target by Octopize (indicative)	CLARITY	ADVANCE
Singling out	-Distance to the closest	> 0.2	0.31	0.93
	-Distance to the closest ratio	> 0.3	0.83	0.84
Linkability	-Column direct match protection	> 50%	83.5%	98.0%
	- Row direct match protection	> 90%	100%	100%
Inference	-Median LC	> 5	4	13
	-Hidden rate	> 90%	84.8%	93.6%
	-Categorical hidden rate	> 90%	97.8%	98.5%



Stanislas DEMUTH – PhD Defense,
Strasbourg, December 18th, 2024

Optimization is possible using the K ; NCP , W parameters

Results: Conservation of univariate distributions



UTILITY (1/5)

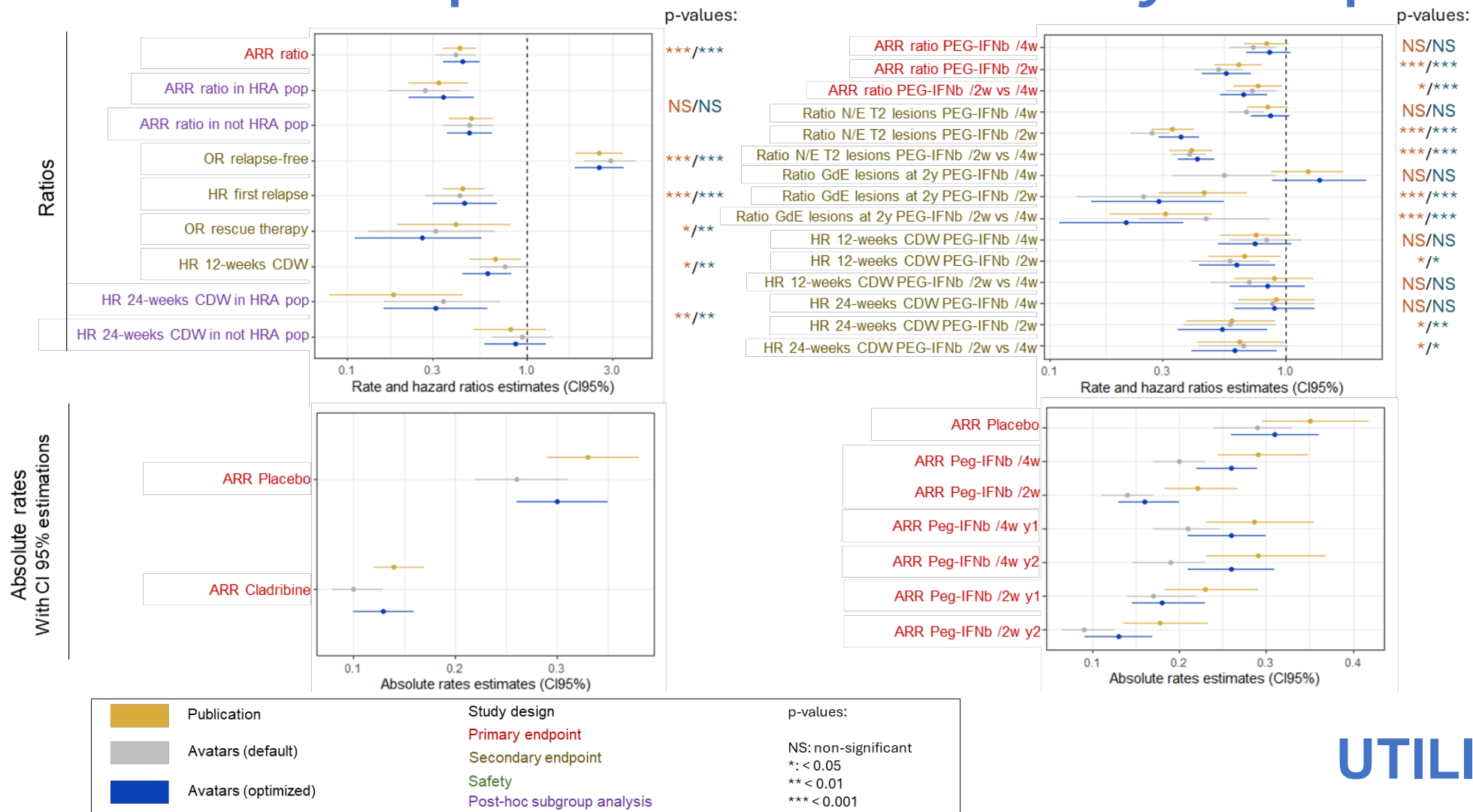
One
Randomised Clinical Trial
ADVANCE

*Stanislas DEMUTH – PhD Defense,
Strasbourg, December 18th, 2024*

CDW: Confirmed disability worsening
GdE: Gadolinium-enhancing
N/E: New or enlarging

Reference data Avatars

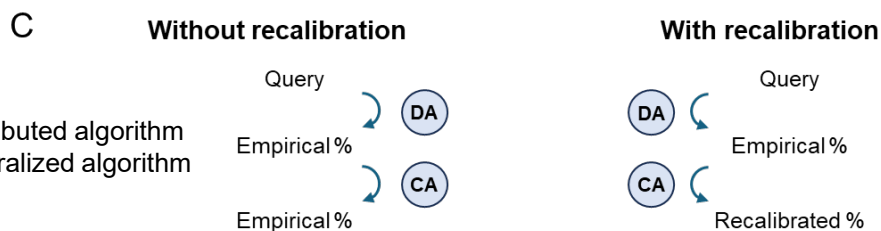
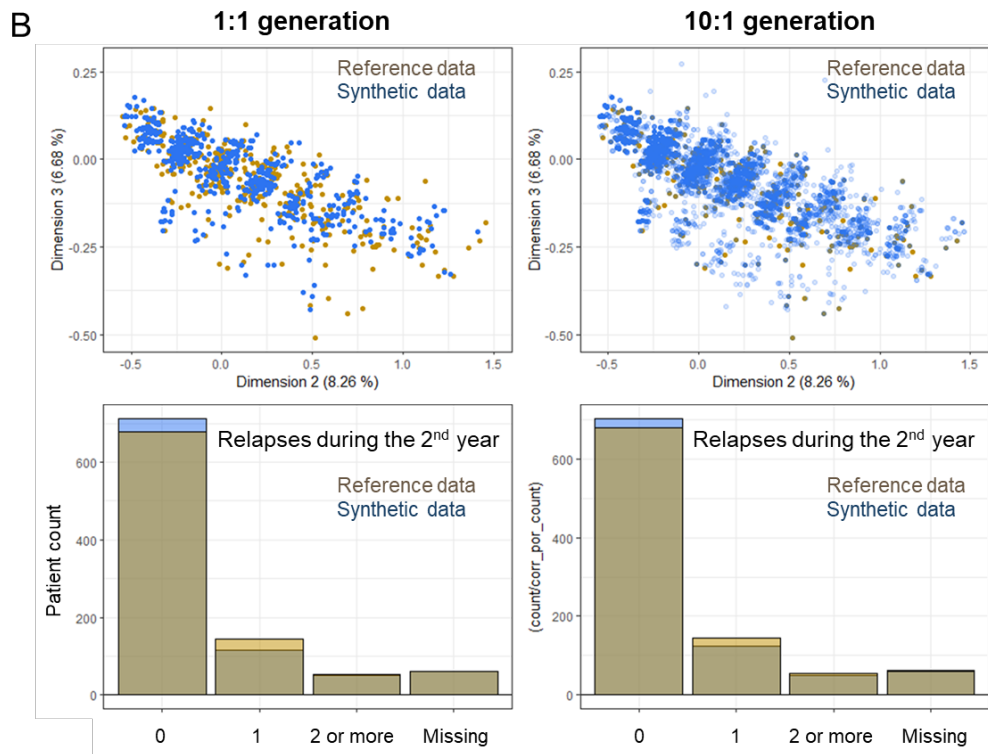
Results: Replication of efficacy endpoints



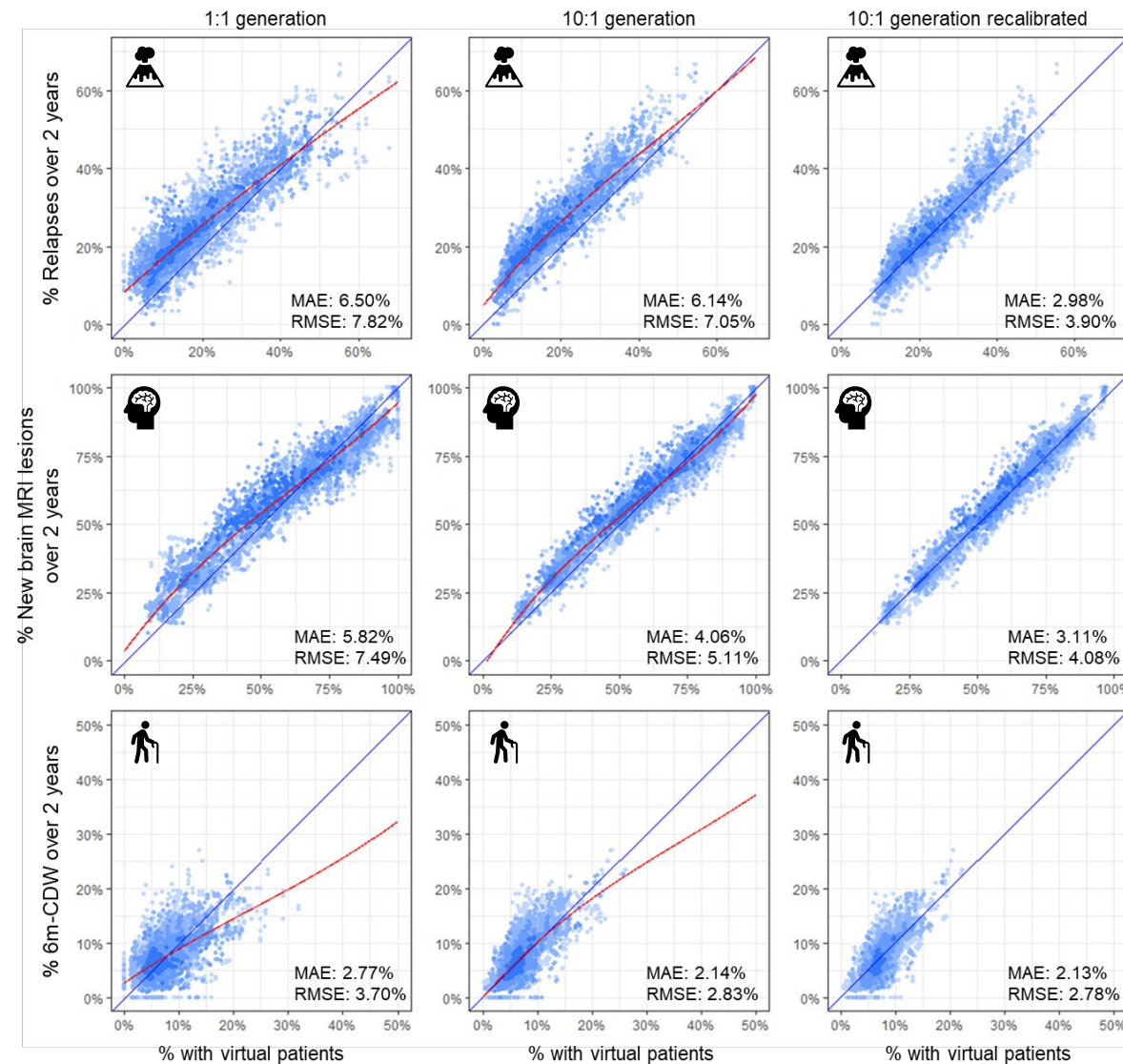
UTILITY (2/5)

Stanislas DEMUTH – PhD Defense,
Strasbourg, December 18th, 2024

Data Augmentation: 1-X, Multi-stratified uses



DA: Distributed algorithm
CA: Centralized algorithm

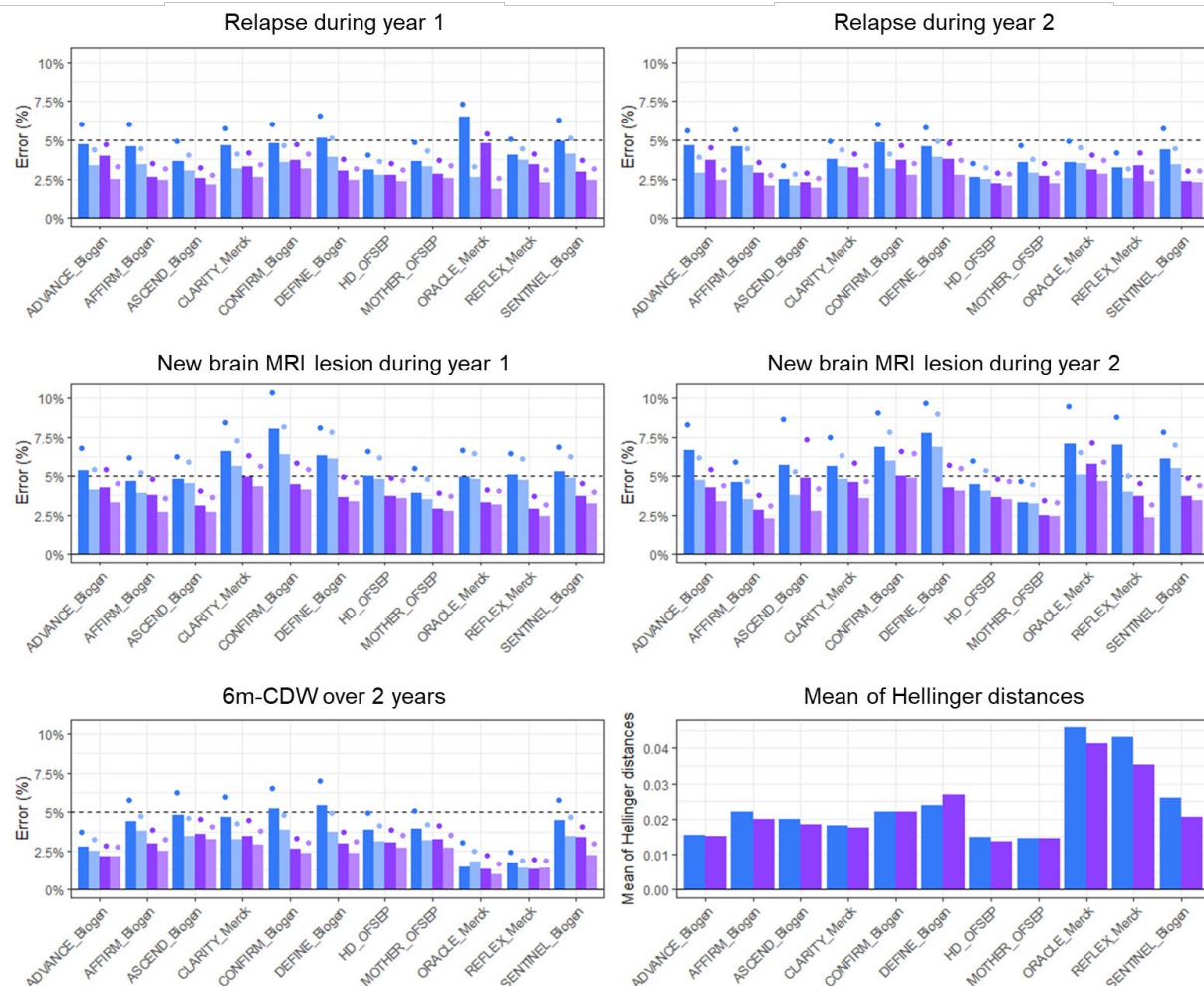


Number of patients: 30 50 100 250

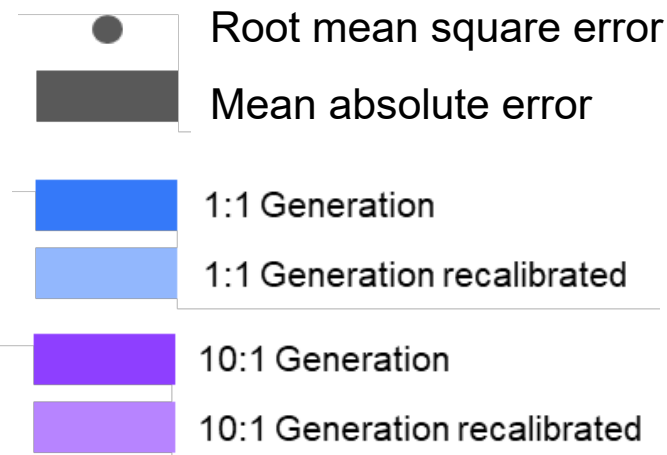
MAE: Mean absolute error
RMSE: Root mean square error

UTILITY (3/5)

Data Augmentation: 1-X, Multi-stratified uses



One
Randomised Clinical Trial
ADVANCE



*Stanislas DEMUTH – PhD Defense,
Strasbourg, December 18th, 2024*

UTILITY (4 /5)

EU GDPR ... *can foster innovation*



- **The term “Personal data” is defined in Art. 4 (1).**
 - Personal data are any information which are related to an identified or identifiable natural person.
- **Article 5 1. on processing of personal data and minimization principle**

1. Personal data shall be:

- (a) ...('lawfulness, fairness and transparency')
- (b) *collected for specified, explicit and legitimate purposes and not further processed (..) statistical purposes shall, in accordance with [Article 89\(1\)](#), not be considered to be incompatible with the initial purposes ('purpose limitation');*
- (c) *adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimization');*
- (d) (...) ('Accuracy')
- (e) *kept in a form which permits identification of data subjects for no longer than is necessary for the purposes (...) statistical purposes in accordance with [Article 89\(1\)](#) (...) ('storage limitation');*
- (f) *processed in a manner that ensures appropriate security of the personal data, (...) , using appropriate technical or organizational measures ('integrity and confidentiality').*

- Right of identifiable person : Trust & transparency **GDPR Article 12** and following ...
- **Article 89** GDPR Safeguards and derogations relating to **processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes**

What is personal health data ?



- **Personal health data** refers to information that relates to an individual's health status, medical history, lifestyle, or genetic makeup. It can be classified based on **nature, destination/use**, and **combination**
- **1. By Nature** data is derived from health information.
 - Ex Medical Data: Diagnoses, medication records, imaging results (e.g., X-rays, MRIs).
 - Ex Biological Data: Blood pressure, heart rate, cholesterol levels.
- **2. By Destination/Use** that makes data relevant for health
 - Ex Lifestyle Data: Diet, physical activity, sleep patterns.
- **3. By Combination** of data from various sources whose combination is relevant for health issues
 - Ex : Multi-Modal Data Fusion: Integration of mobility decay, and walking images

-Omics data is personal health data



- **GDRP is both a constraint and an opportunity**
 - Reassess how we manage biomedical data from patients
- **Genetic data is both pioneer and large scale**
 - Quantitative leap forward since the late 90's
 - 1.5 millions genotype (A/T/C/G, A/T/C/G)
 - ~ 80 euros /patient

Genetic data is not anonymous

- **Pseudonymous data** (ARTICLE 29 DATA PROTECTION WORKING PARTY)
 - (i) *is it still possible to single out an individual ? Yes...*
 - (ii) *is it still possible to link records relating to an individual ? Yes...*
 - (iii) *can information be inferred concerning an individual ? Yes...*

Scientific context : (Health) Data Sharing

1- Research cohort/ Data Resources

- Historical Ex IHIW
- Ex: 1000 Genome, HLA
- Ex: UK Biobank etc

2- Biomedical data Warehouse (BDW)

- Data Clinic – Karakachoff et al 2024 JMIR

• Open Science Data Repository (OSDR)

- → *Data sharing drives innovation and knowledge transfer*

Why Sharing ? What controls over data usages (claims) ? What about Privacy ? What are the levels of sensitivity that should be addressed ?

	N	Structured	Research-ready
1-Research	++	+	+++
2- BDW	+++	+ /-	-
3- OSDR	-	+++	++

Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight

- Karakachoff M, Goronflot T, Coudol S, Toublant D, Bazoge A, Constant Dit Beaufils P, Varey E, Leux C, Mauduit N, Wargny M, Gourraud P JMIR Med Inform 2024;12:e50194

<https://medinform.jmir.org/2024/1/e50194>

“ We present the actions and challenges regarding organizational changes, technical architecture, and shared governance that took place to develop the Nantes BDW. We describe the process to access clinical contents, give details about patient data protection, and use examples to illustrate merging clinical insights.”



JMIR Publications
Advancing Digital Health & Open Science

25
years

DOI: 10.2196/50194

Implementation Report

Implementing a Biomedical Data Warehouse From Blueprint to Bedside in a Regional French University Hospital Setting: Unveiling Processes, Overcoming Challenges, and Extracting Clinical Insight

Matilde Karakachoff¹, MSc; Thomas Goronflot¹, MSc; Sandrine Coudol¹, MSc; Delphine Toublant^{1,2}, MSc; Adrien Bazoge^{1,3}, MSc, PhD; Pacôme Constant Dit Beaufils^{1,4}, MD; Emilie Varey^{1,5}, MSc; Christophe Leux⁶, MD, PhD; Nicolas Mauduit⁶, MD; Matthieu Wargny¹, MD, PhD; Pierre-Antoine Gourraud^{1,7}, MPH, PhD

¹Centre d'Investigation Clinique 1413, INSERM, Clinique des données, Pôle Hospitalo-Universitaire 11: Santé Publique, Centre Hospitalier Universitaire Nantes, Nantes Université, Nantes, France

²IT Services, Centre Hospitalier Universitaire Nantes, Nantes Université, Nantes, France

³Unité Mixte de Recherche 6004, Laboratoire des Sciences du Numérique de Nantes, Centre National de Recherche Scientifique, École Centrale Nantes, Nantes Université, Nantes, France

⁴Institut du thorax, Service de neuroradiologie diagnostique et interventionnelle, Centre Hospitalier Universitaire Nantes, Nantes Université, Nantes, France

⁵Direction de la Recherche et de l'Innovation, Centre Hospitalier Universitaire Nantes, Nantes Université, Nantes, France

⁶Service d'information médicale, Centre Hospitalier Universitaire Nantes, Nantes Université, Nantes, France

⁷INSERM Center for Research in Transplantation and Translational Immunology, Nantes Université, Nantes, France

Corresponding Author:

Pierre-Antoine Gourraud, MPH, PhD

Centre d'Investigation Clinique 1413, INSERM, Clinique des données

Pôle Hospitalo-Universitaire 11: Santé Publique, Centre Hospitalier Universitaire Nantes, Nantes Université

30 bd Jean Monnet

Nantes, 44093

France

Phone: 33 2 44768234

Email: pierre-antoine.gourraud@univ-nantes.fr

Abstract

Background: Biomedical data warehouses (BDWs) have become an essential tool to facilitate the reuse of health data for both research and decisional applications. Beyond technical issues, the implementation of BDWs requires strong institutional data governance and operational knowledge of the European and national legal framework for the management of research data access and use.

Objective: In this paper, we describe the compound process of implementation and the contents of a regional university hospital BDW.

Methods: We present the actions and challenges regarding organizational changes, technical architecture, and shared governance that took place to develop the Nantes BDW. We describe the process to access clinical contents, give details about patient data protection, and use examples to illustrate merging clinical insights.

Implementation (Results): More than 68 million textual documents and 543 million pieces of coded information concerning approximately 1.5 million patients admitted to CHUN between 2002 and 2022 can be queried and transformed to be made available to investigators. Since its creation in 2018, 269 projects have benefited from the Nantes BDW. Access to data is organized according to data use and regulatory requirements.

Conclusions: Data use is entirely determined by the scientific question posed. It is the vector of legitimacy of data access for secondary use. Enabling access to a BDW is a game changer for research and all operational situations in need of data. Finally, data governance must prevail over technical issues in institution data strategy vis-à-vis care professionals and patients alike.

JMIR Med Inform 2024;12:e50194; doi: [10.2196/50194](https://doi.org/10.2196/50194)

<https://medinform.jmir.org/2024/1/e50194>

JMIR Med Inform 2024 | vol. 12 | e50194 | p. 1
(page number not for citation purposes)

Data Governance : Global Access to multiple data resources ...

		LEGAL RESPONSIBILITY OF THE DATA (NANTES UNIVERSITÉ MEDICAL CENTER) ?	
		YES	NO
INTENTIONALITY <i>Were data recorded for research purposes ?</i>	NO	Biomedical Data Warehouse of the Nantes Université Medical Center 	National Medico-Administrative Data <i>(Reimbursement national database)</i> 
	YES	Hospital research Cohorts BRUGADA, ICAN, EASY, VISIOCORT, EXAN, COVER, VALIDate, CoHPT, IT-DIAB, CORONADO...  	National or international Databases made available for research 

The Data Clinic : *a house to care for data*

- Hospital Unit organized as such
 - Formerly a epidemiological and clinical research methodological support group
- Approved by the French National Data protection Agency in July 2018
- 201 projects in 2023
 - 220 projects in 2022
- Publications 2019-2023 : **143 publications**



Governance Matrix		Legal Resp. (CHUN)	
		Yes	No
Research Intentionality	No	Biomedical Data Warehouse	SNDS
	Yes	Registry Cohorts	Nat. or Int. Databases



**> 3 millions
Patients**



**>700 millions
Structured
Data**



**> 100 millions
Textual
Documents**

National/EU Regulation GDPR

... *can foster innovation*



https://www.cnil.fr/sites/default/files/atoms/files/referentiel_entrepot.pdf

Page 14 : Exportation de données hors de l'entrepôt et hors des espaces de travail

SEC-EXP-1 A l'exception des données relatives aux procédures de ré-identification SEC-REI-1 à SEC-REI-3, **seuls des jeux de données anonymes peuvent faire l'objet d'une exportation hors de l'entrepôt ou d'un espace de travail**. Le processus d'anonymisation doit produire un jeu de données conforme **aux trois critères définis par l'avis du G29 n° 05/2014** ou à tout avis ultérieur du CEPD relatif à l'anonymisation. **Cette conformité doit être documentée et démontrable**. À défaut, si ces trois critères ne peuvent être réunis, une étude des risques de ré-identification devra être menée et documentée.

SEC-EXP-2 Les exports de données doivent être soumis à **la validation préalable d'un responsable** afin d'en avaliser le principe, notamment au regard de l'exigence SEC-EXP-1.

SEC-EXP-3 Les exports doivent faire l'objet d'une surveillance automatique ou manuelle par un opérateur spécialisé afin d'en vérifier le caractère anonyme. Dans le cas où cette surveillance est automatique, tout export identifié comme non conforme doit faire l'objet d'une remontée d'alerte et d'une mise en quarantaine dans l'entrepôt, puis doit être vérifié manuellement par un responsable spécifiquement formé et spécifiquement habilité.

SEC-EXP-4 Les systèmes mis en place dans l'entrepôt relatifs à la production d'indicateurs et au pilotage stratégique de l'activité d'un établissement de santé ne doivent permettre que des restitutions anonymes, y compris en tenant compte des fonctionnalités de filtrage et de sélection de ces restitutions. Ce processus de restitution doit être conforme aux trois critères définis par l'avis du G29 n° 05/2014 ou à tout avis ultérieur du CEPD relatif à l'anonymisation. Cette conformité doit être documentée. À défaut, si ces trois critères ne peuvent être réunis, **une étude des risques de ré-identification devra être menée et documentée**.