Nantes Université

Génération de données synthétiques centrées sur le patient, aucune raison de risquer la ré identification dans l'analyse des données biomédicales

Amphi Schwarz, IMT 11:30 - 12:00

Pr Pierre-Antoine GOURRAUD, PhD MPH

Vendredi 9 Février 24, Toulouse

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

Nantes Université

- **COI :**

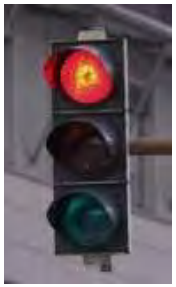
PA Gourraud est le fondateur (2008) (www.methodomics.com) et le co-fondateur de Big data Santé (2018) . Il est consultant et/ou intervenant pour de grandes entreprises pharmaceutiques ou de dispositifs médicaux. Ses activités sont toutes traitées par une contractualisation universitaire ou hospitalière*(AstraZeneca, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Janssen, IAGE, Lek, Methodomics, Merck, Mérieux,Octopize, Sanofi-Genzyme)*. 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

- ***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, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Janssen, IAGE, Lek, Methodomics, Merck, Mérieux, Octopize, Sanofi-Genzyme). 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.

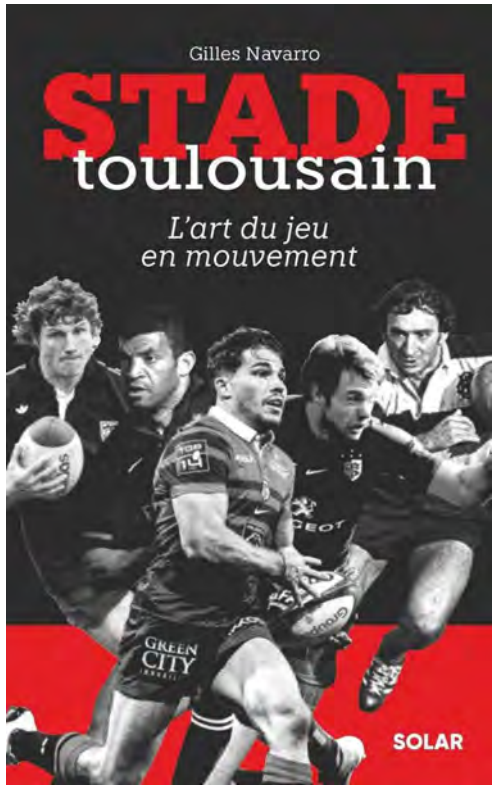
Nantes Université

Le problème de données de Santé ….



La circulation des données …

La solution du jeu à la Toulousaine… ***une combinaison de mouvements et de passes, mettant la vitesse d'exécution, Elle permet de créer des opportunités d'attaques.***



Les Données Personnelles de santé NE CIRCULENT PAS
… Car soumises :
+ au RGPD
+ au référentiel EDS CNIL,
+ à la pression des patients,
+ à la pression des soignants,
+ à la pression des établissements.

« 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. » Référentiel CNIL EDS

CNIL. COMMISSION NATIONALE INFORMATIQUE & LIBERTÉS

Pôle Santé UFR Médecine · DELPHI AI & Health data · France 2030 · anr · Région Pays de la Loire · Nantes Métropole · Centre Hospitalier Universitaire de Nantes · Nantes Université · Inserm Institut national de la santé et de la recherche médicale

Un contexte:

La Clinique des données du CHU de NANTES, EDS, faciliter les usages secondaire de la donnée issue du soin

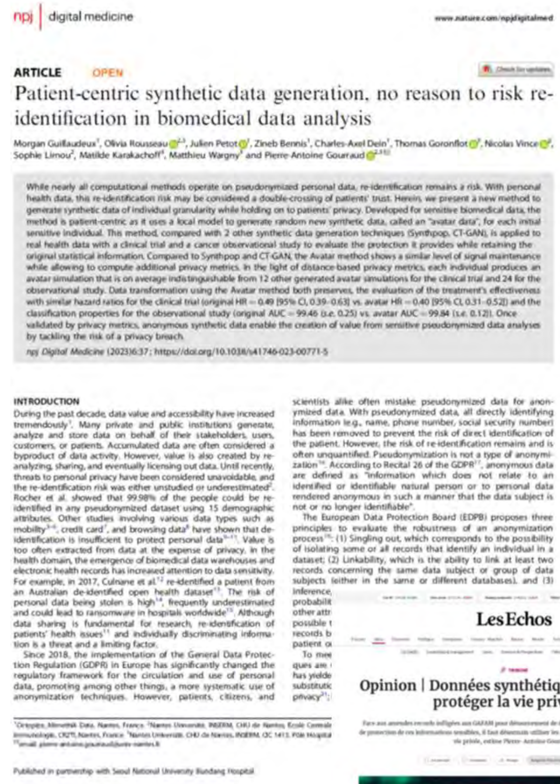
Mettre en qualité la donnée de santé pour la rendre réutilisable

La solution : *Simuler des données anonymes à (très haute) valeur informative*

Comment faire des données anonymes démontrables ?

Guillaudeux et al 2023 **Nature Digital Medicine**

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



npj | digital medicine www.nature.com/npjdigitalmed

ARTICLE OPEN

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

Morgan Guillaudeux[1], Olivia Rousseau[2,3], Julien Petot[1], Zineb Bennis[1], Charles-Axel Dein[1], Thomas Goronflot[3], Nicolas Vince[2], Sophie Limou[2], Matilde Karakachoff[4], Matthieu Wargny[3] and Pierre-Antoine Gourraud[2,3,5]

While nearly all computational methods operate on pseudonymized personal data, re-identification remains a risk. With personal health data, this re-identification risk may be considered a double-crossing of patients' trust. Herein, we present a new method to generate synthetic data of individual granularity while holding on to patients' privacy. Developed for sensitive biomedical data, the method is patient-centric as it uses a local model to generate random new synthetic data, called an "avatar data", for each initial sensitive individual. This method, compared with 2 other synthetic data generation techniques (Synthpop, CT-GAN), is applied to real health data with a clinical trial and a cancer observational study to evaluate the protection it provides while retaining the original statistical information. Compared to Synthpop and CT-GAN, the Avatar method shows a similar level of signal maintenance while allowing to compute additional privacy metrics. In the light of distance-based privacy metrics, each individual produces an avatar simulation that is on average indistinguishable from 12 other generated avatar simulations for the clinical trial and 24 for the observational study. Data transformation using the Avatar method both preserves, the evaluation of the treatment's effectiveness with similar hazard ratios for the clinical trial (original HR = 0.49 [95% CI, 0.39–0.63] vs. avatar HR = 0.40 [95% CI, 0.31–0.52]) and the classification properties for the observational study (original AUC = 99.46 (s.e. 0.25) vs. avatar AUC = 99.84 (s.e. 0.12)). Once validated by privacy metrics, anonymous synthetic data enable the creation of value from sensitive pseudonymized data analyses by tackling the risk of a privacy breach.

npj Digital Medicine (2023)6:37 ; https://doi.org/10.1038/s41746-023-00771-5

INTRODUCTION

npj | Digital Medicine

Nantes Université

Triple Actualité des données Synthétiques

1. Actualité éthique : Avis *« Plateformes de données de santé : enjeux d'éthique »*

- **Avis 143** : Comité consultatif national d'éthique (CCNE)
- **Avis 5** : Comité national pilote d'éthique du numérique (CNPEN)
- Page 31 - Recommandation 8
  - *https://www.ccne-ethique.fr/publications/avis-143-du-ccne-*

2. Actualité réglementaire : RGPD pour les usages données personnelles (de santé)

- **(Vs.) Données publiques (Open Source) - Confidentialité**
- **Contrôle des finalités, l'identité des usagers des données,**
- **Référentiel EDS** publié par la CNIL : **« Données Anonymes »**

3. Actualité scientifique : Usages des données Synthétiques

- Lauréat de l'**Appel à projet Entrepôt de Données de Santé** (EDS) : « ODH 2.0 »
  - Projet technique Données Synthétiques Anonymes Avatars avec Octopize
- **"Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis"**

Guillaudeux et al 2023 Nature Digital Medicine
https://www.nature.com/articles/s41746-023-00771-5

Nantes Université

La Méthode

La Clinique des données du CHU de Nantes favorise les usages secondaires des données de santé.


npj | digital medicine

ARTICLE OPEN

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

Comment faire des données anonymes démontrables ?

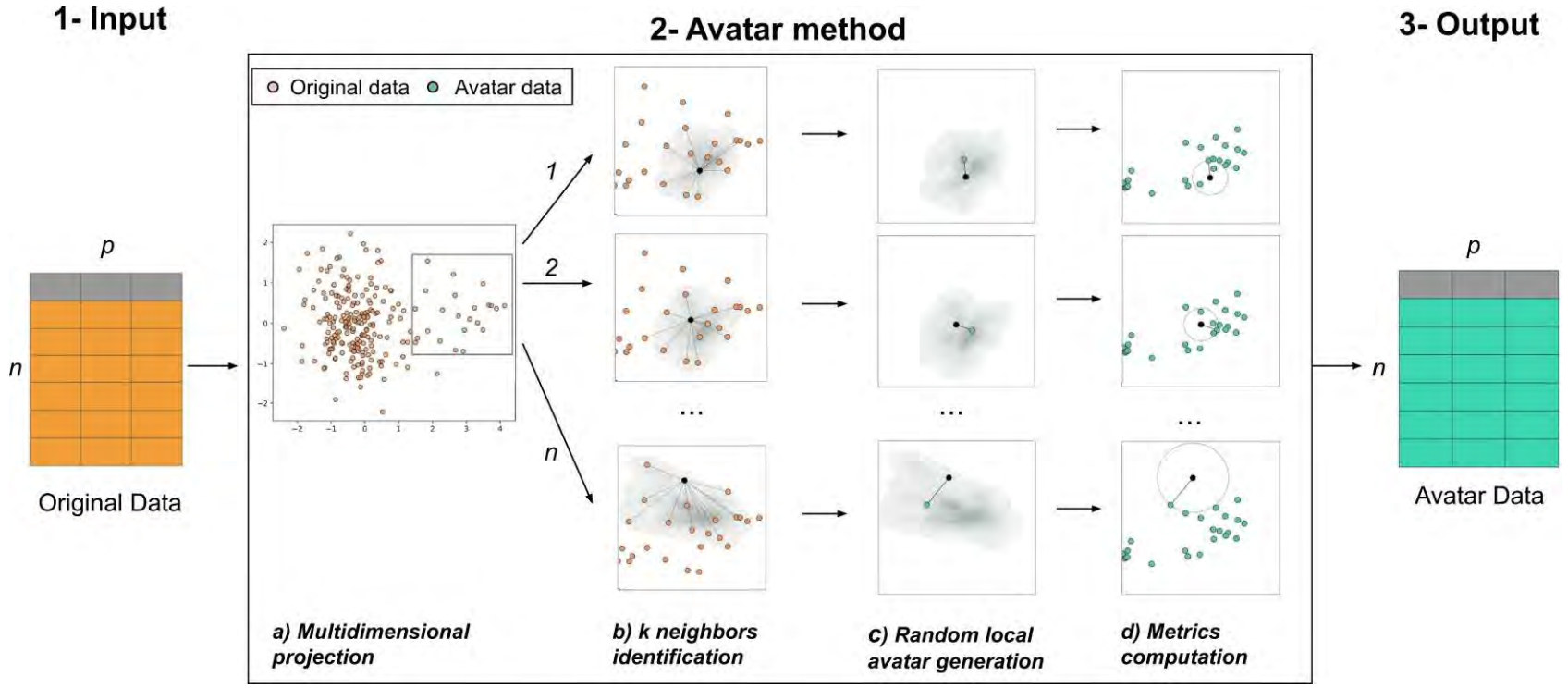
Guillaudeux et al 2023 **Nature Digital Medicine**

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

npj | Digital Medicine

Methodes: Données Synthétiques anonymes

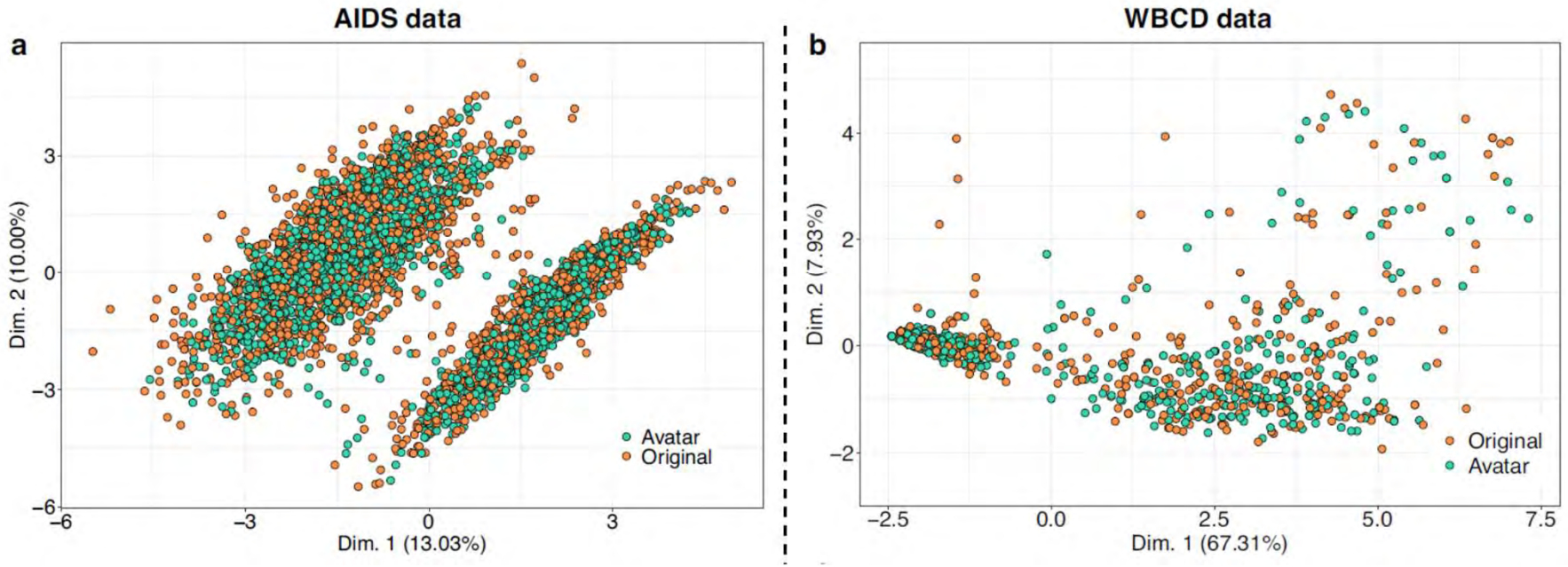
- **Comment ça marche ?** Un modèle privé non-paramétrique à usage unique



Guillaudeux et al 2023 Nature Digital Medicine - https://www.nature.com/articles/s41746-023-00771-5 Figure 5

Results 1: Datasets conservations

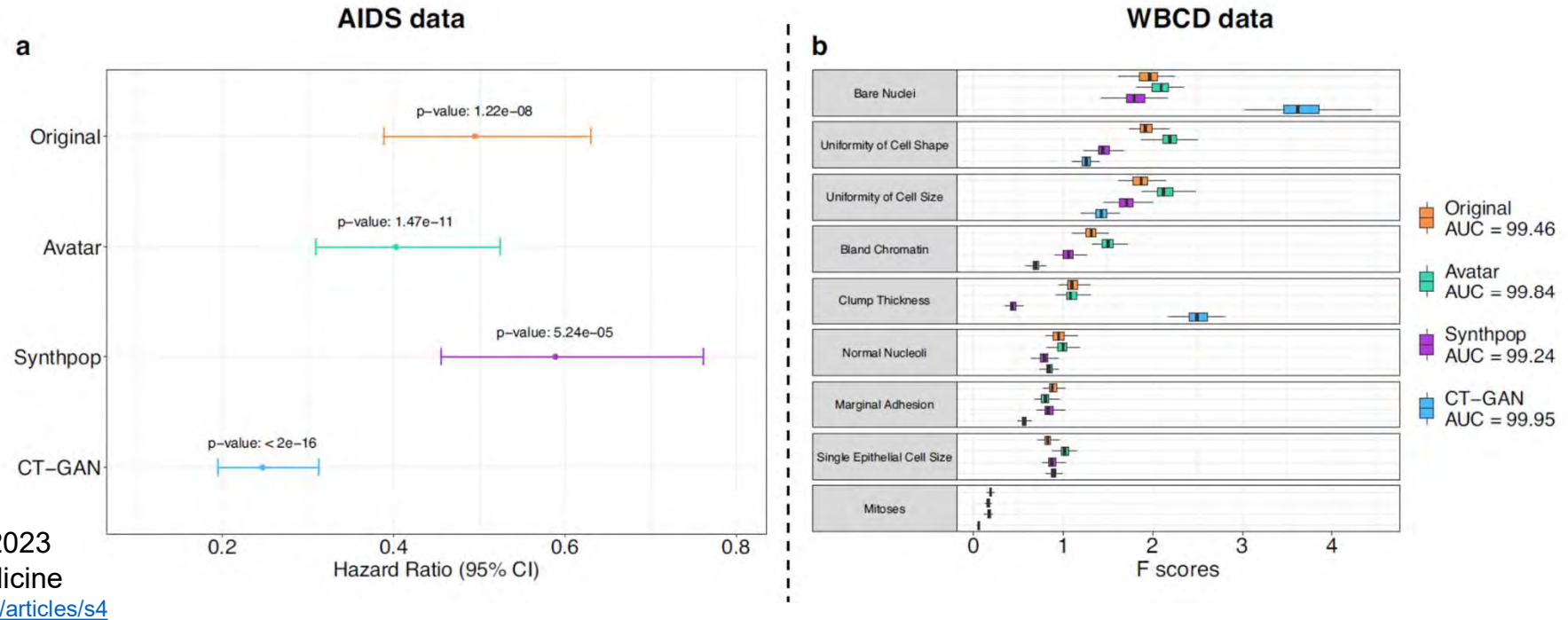
- Représentation multidimensionnelles semblables



Guillaudeux et al 2023
Nature Digital Medicine
https://www.nature.com/articles/s41746-023-00771-5 Figure 1 A&B

Results 2: Statistics conservations

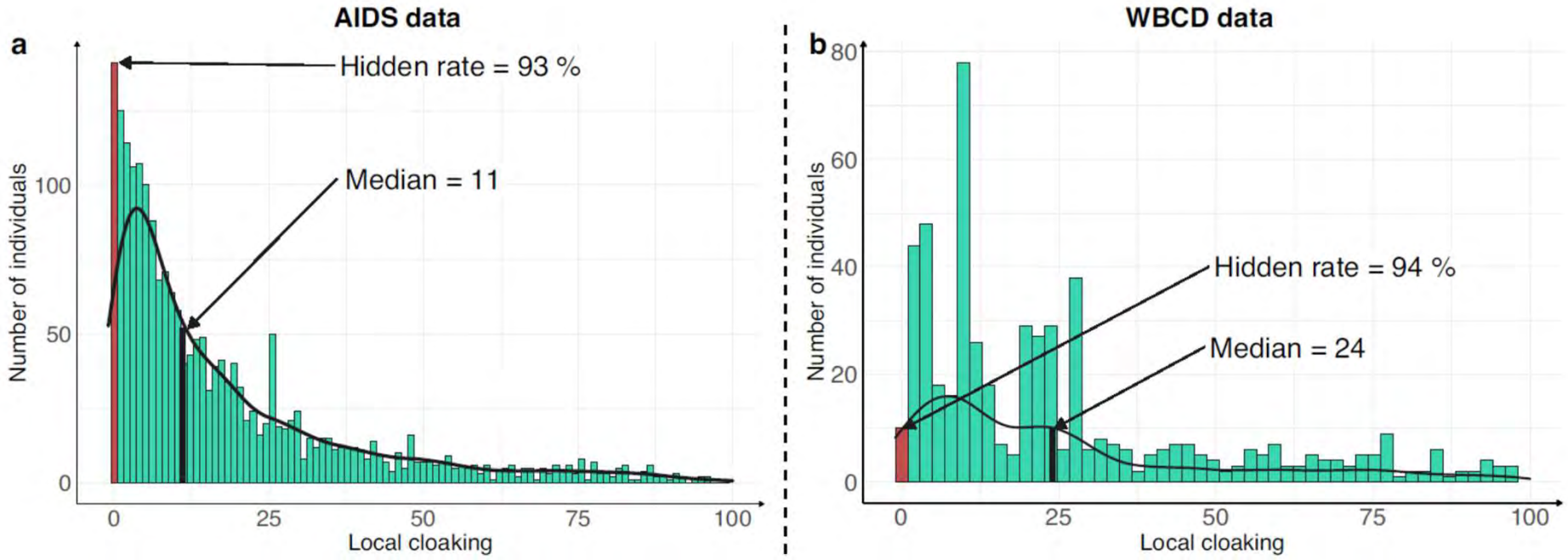
- Résultats d'ensemble similaires à l'analyse principale (si existante) + autres méthodes



Guillaudeux et al 2023
Nature Digital Medicine
https://www.nature.com/articles/s41746-023-00771-5 Figure 2 A&B

Resultats 3 : Est ce que la valeur statistique des données compte vraiment en premier ?

- Les métriques qui attestent que les Avatars sont des données synthétiques anonymes au sens CNIL - c'est-à-dire au sens démontrable.



Guillaudeux et al 2023 Nature Digital Medicine https://www.nature.com/articles/s41746-023-00771-5 Figure 3A & 3B voir 3C & 3D pour plus de garanties

Nantes Université

Application dans le contexte des entrepôts des données de santé

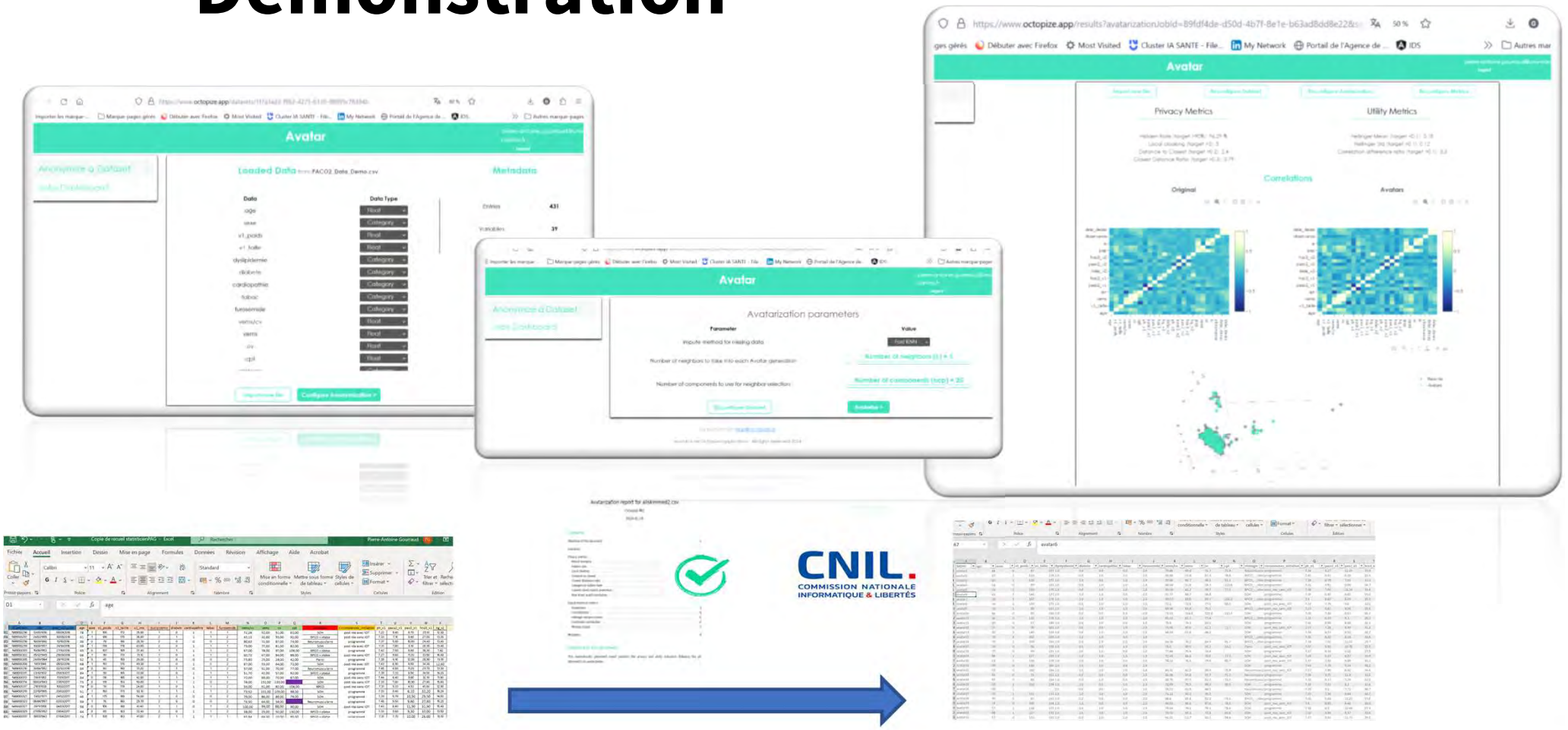
La Clinique des données du CHU de Nantes favorise les usages secondaires des données de santé.

Un cas d'usage : Publication & Science Ouverte

- ***Situation:*** le Pr FX B et le DR S J demandent à la Clinique des données de l'aide pour répondre aux reviewers dernier obstacle pour la publication.
  - ***"Does $PaCO_2$ correction have an impact on survival of patients with chronic respiratory failure and long-term non-invasive ventilation?"***
- **Commentaire 1 – rev 1 :** Etude avec manque de puissance statistique, hétérogénité Clinique, possibilité de partager les données (en vue de méta analyse) ?
  - *Données pseudonymisées – Contrôle des usages*
- **Solution :** Données synthétiques anonymes
  - Application du référentiel "EDS"
- **Autre exemple :** JNNP en open source (licence MIT)
  - https://jnnp.bmj.com/content/92/2/122
  - https://github.com/ICAN-aneurysms/RIA-predict/tree/master/notebooks

Nantes Université

Démonstration



Nantes Université

Application dans le contexte des entrepôts des données de santé :

Projet RHU Médecine de précision pour les maladies chroniques

Exemples en maladie chronique : Sclérose en plaques et Transplantation

RHU KTD-Innov - Ref projet: IA-17-RHUS-0010 : Ce travail bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du PIA, portant la référence IA-17-RHUS-0010

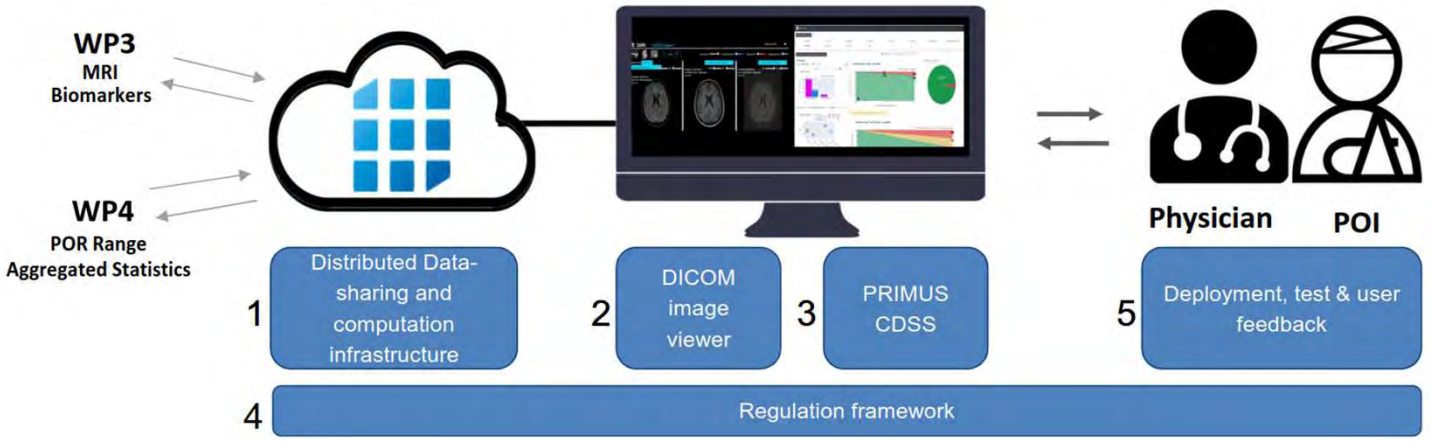
RHU PRIMUS - Ref projet: ANR-21-RHUS-0014 : Ce travail bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du 3e PIA, intégré à France 2030 portant la référence ANR-21-RHUS-0014

Nantes Université

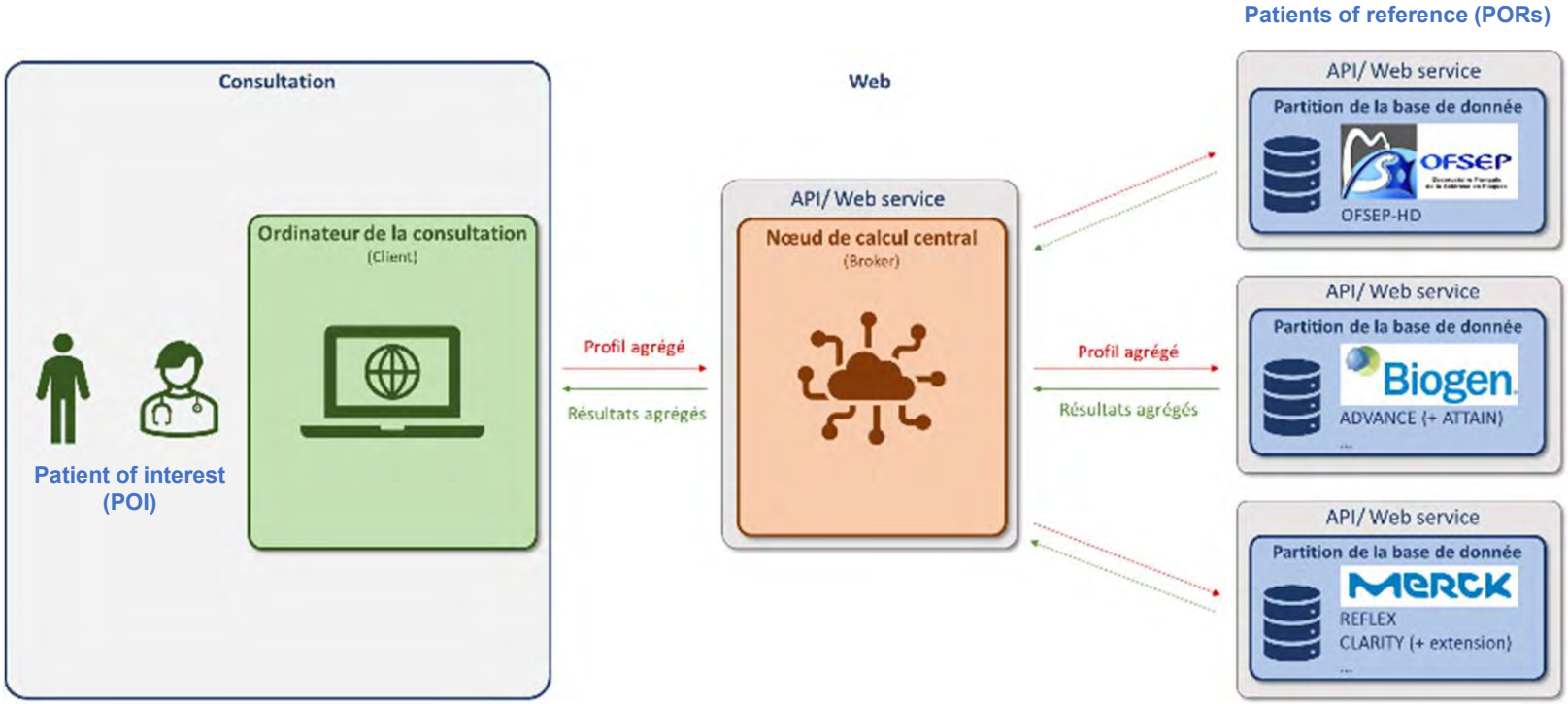
Projet de RHU Primus en Sclérose en plaques

Outils d'aide à la décision médicale (**CDSS**) qui permet au neurologue, lors d'une consultation avec son patient (Patient Of Interest – **POI**) de visualiser (**viewer**) la projection de l'évolution de la maladie du POI par rapport à son imagerie (**IRM**) et des algorithmes collaboratifs (**IA**) qui vont sélectionner des patients de références (**POR**) dans une base de données de références (**BD Ker PRIMUS**) construite à partir des essais cliniques de l'industrie et des données de vie réelles de plus de 12 000 individus en accord avec la réglementation (**RGPD**) et le Medical Device Regulation (**MDR**). Testé dans un essai clinique en prévision d'un **marquage CE**.

« Le bon traitement pour la bonne personne au bon moment et prévenir la survenue de handicaps »



Projet de RHU Primus en Sclérose en plaques

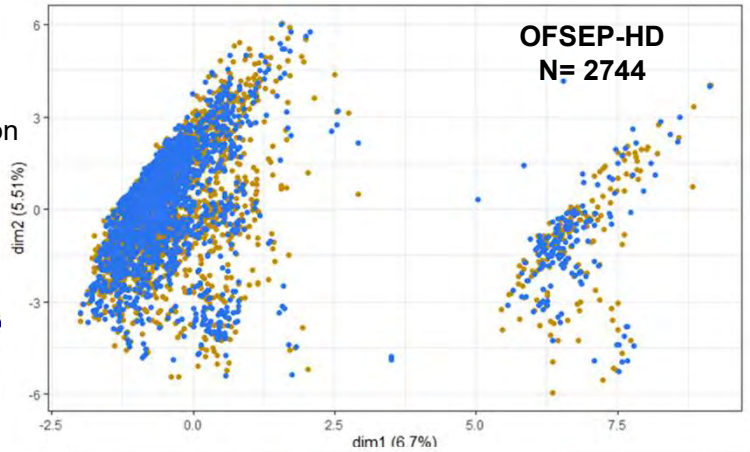


➔ Base de données de référence distribuée: 1 partition = 1 propriétaire de données

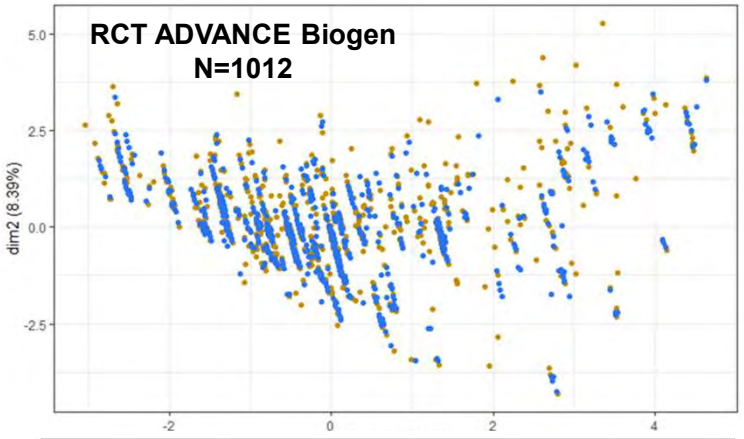
Le logiciel d'aide à la décision du RHU PRIMUS, Dr. Stanislas DEMUTH, laboratoires des Prof. Pierre-Antoine GOURRAUD & Prof. Jérôme DE SEZE, Assises de l'OFSEP 2023

RHU PRIMUS: 4 reference synthetic Datasets
MS Patients – Strategic alliance

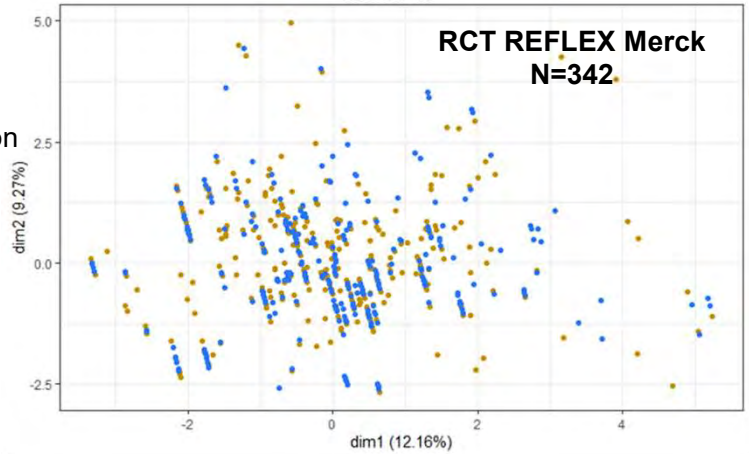
K = 5
NCP = 56 with categories reduction
HR = 83.0 %
Median LC = 3
HM = 0.04



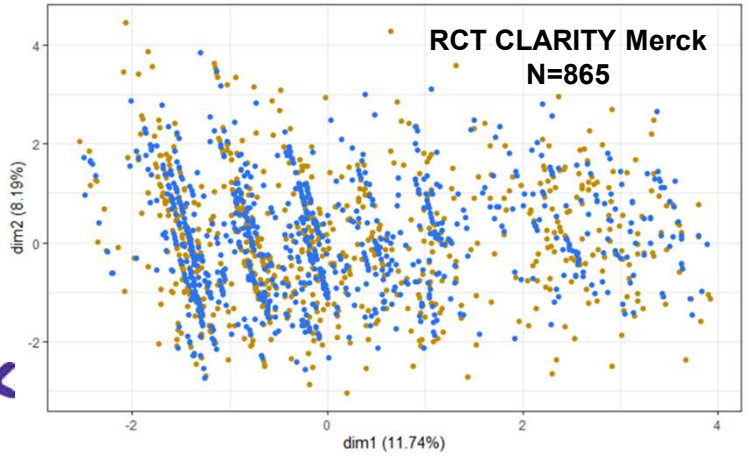
K = 5
NCP = 18 with categories reduction
HR = 86.7 %
Median LC = 3
HM = 0.02



K = 5
NCP = 16 with categories reduction
HR = 82.1 %
Median LC = 2
HM = 0.04



K = 5
NCP = 19 with categories reduction
HR = 82.4 %
Median LC = 2
HM = 0.04

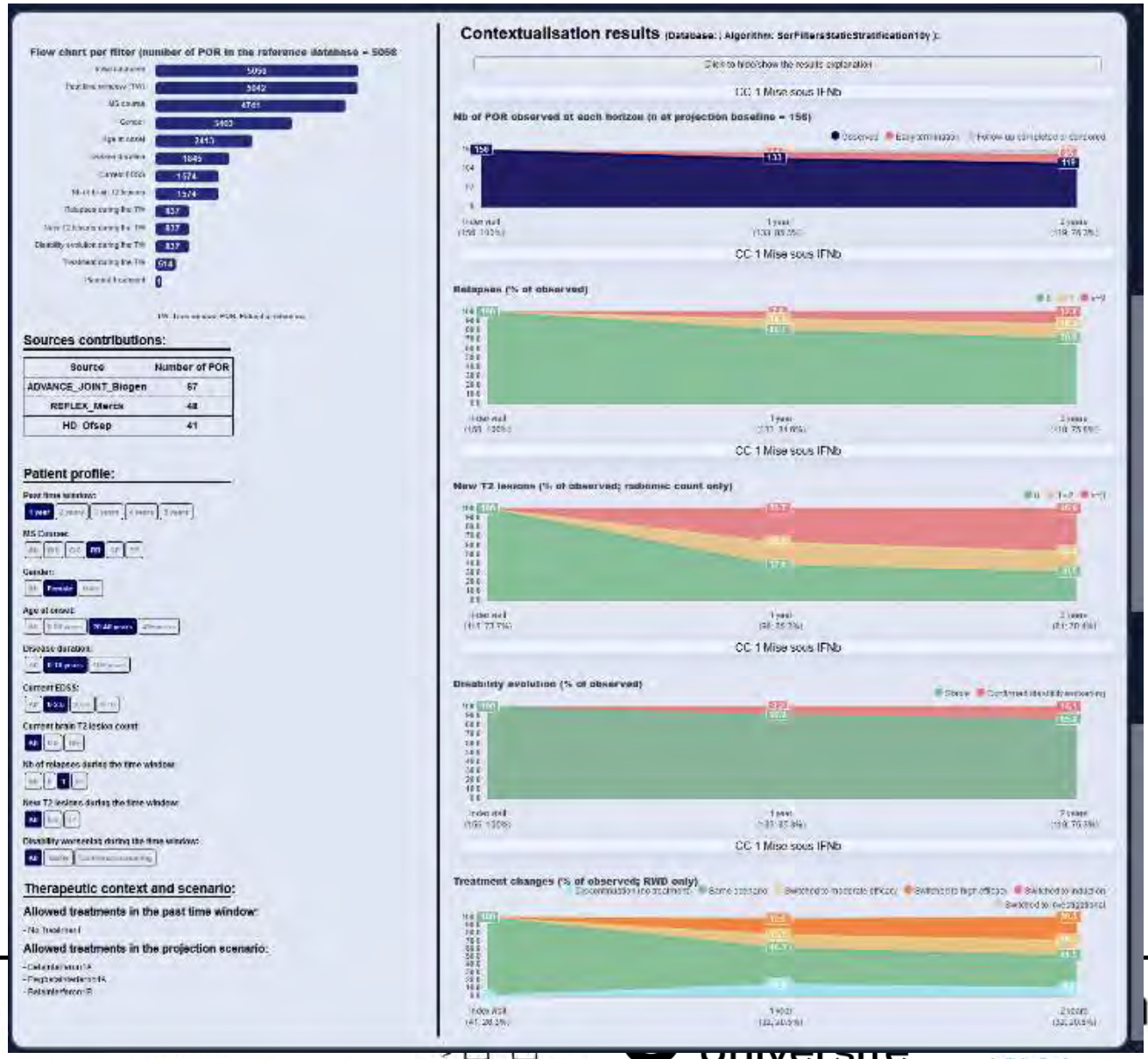


Nantes Université

PRIMUS Alpha : Projection de toute la population de référence

Cas clinique 1 :
♀ 26a, première poussée, McDo+

Mise sous IFN béta



Le logiciel d'aide à la décision du RHU PRIMUS, Dr. Stanislas DEMUTH, laboratoires des Prof. Pierre-Antoine GOURRAUD & Prof. Jérôme DE SEZE, Assises de l'OFSEP 2023

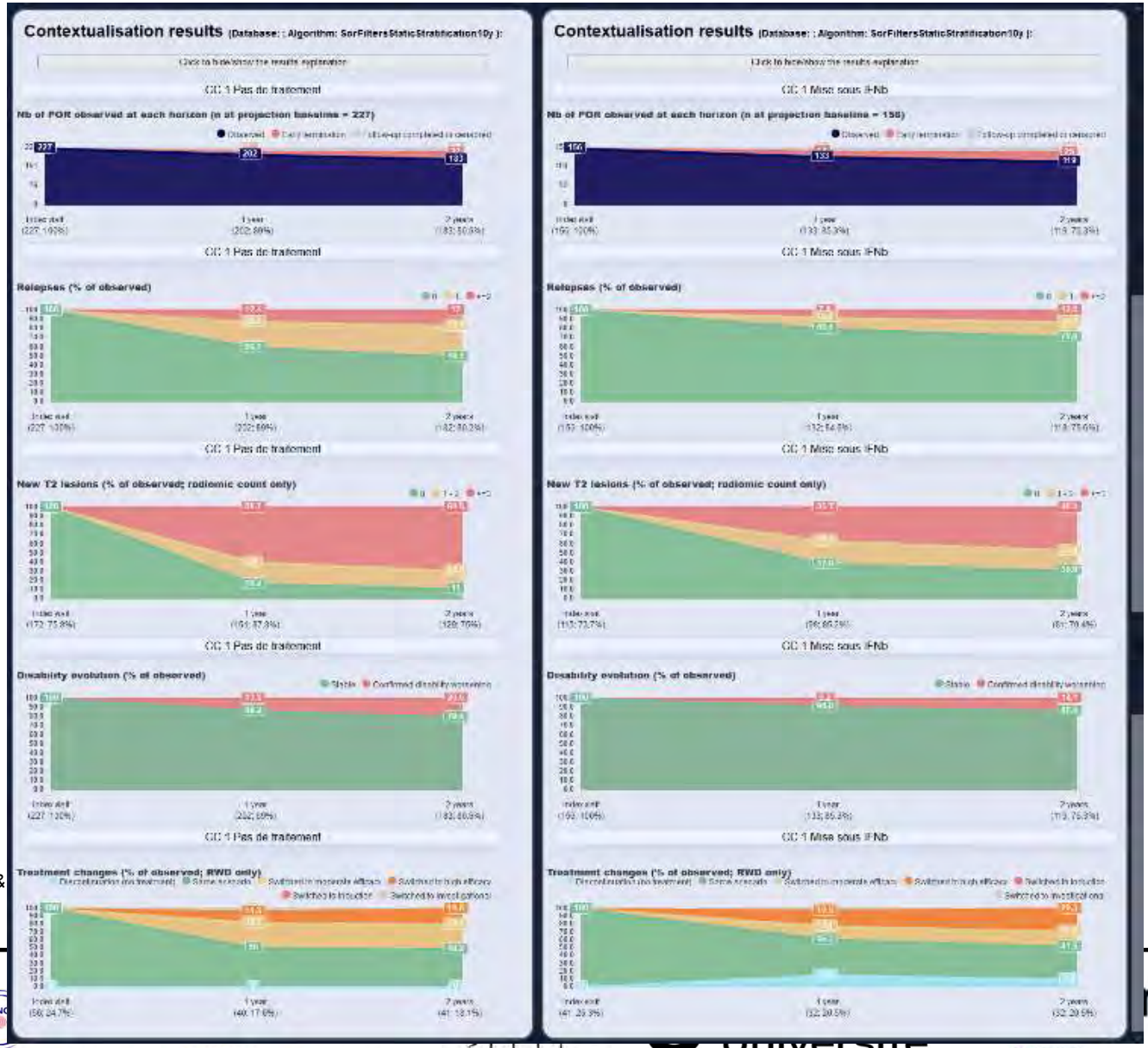
Pôle Santé UFR Médecine
DELPHI Deep Learning Proposal for Healthcare & Innovation — AI & Health data
CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES
Institut national de la santé et de la recherche médicale

Nantes Université

PRIMUS Alpha : Projection de toute la population de référence

Cas clinique 1 :
♀ 26a, première poussée, McDo+

Mise sous IFN béta



Le logiciel d'aide à la décision du RHU PRIMUS, Dr. Stanislas DEMUTH, laboratoires des Prof. Pierre-Antoine GOURRAUD & Prof. Jérôme DE SEZE, Assises de l'OFSEP 2023

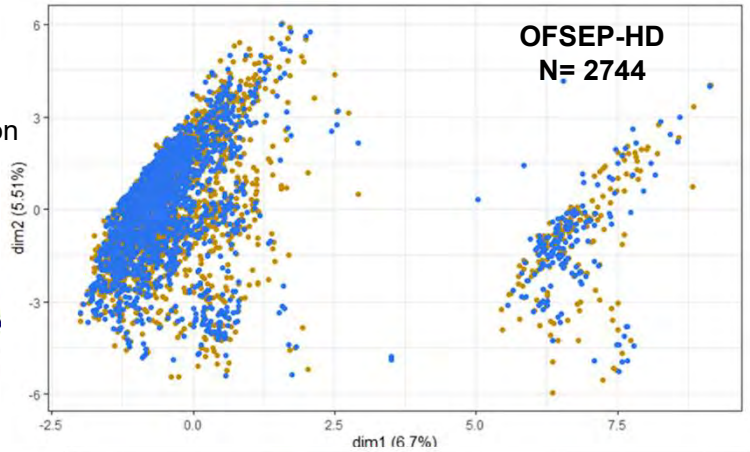
Nantes Université

RHU PRIMUS: 4 reference synthetic Datasets
MS Patients – Strategic alliance

OCTOPIZE · CNIL. COMMISSION NATIONALE INFORMATIQUE & LIBERTÉS · RGPD

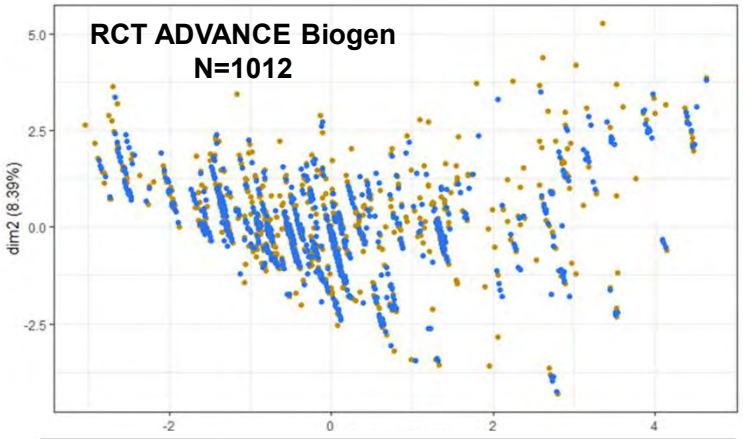
K = 5
NCP = 56 with categories reduction
HR = 83.0 %
Median LC = 3
HM = 0.04

OFSEP Observatoire Français de la Sclérose en Plaques



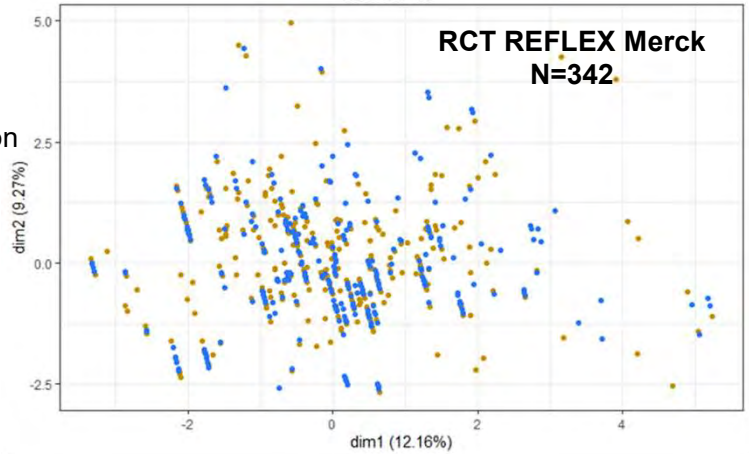
K = 5
NCP = 18 with categories reduction
HR = 86.7 %
Median LC = 3
HM = 0.02

Biogen



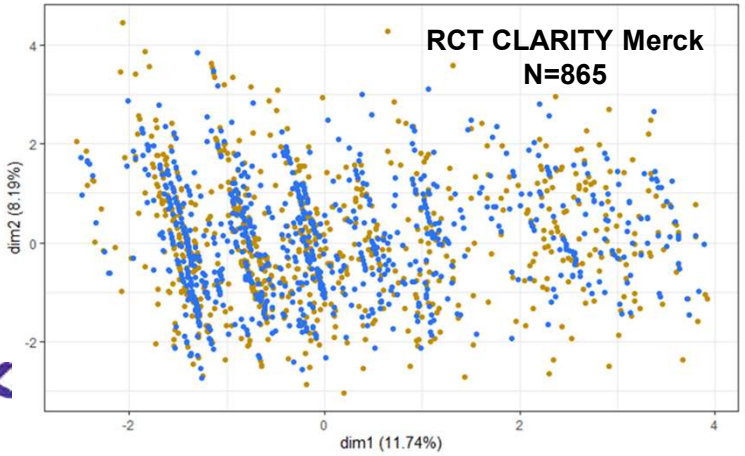
K = 5
NCP = 16 with categories reduction
HR = 82.1 %
Median LC = 2
HM = 0.04

MERCK



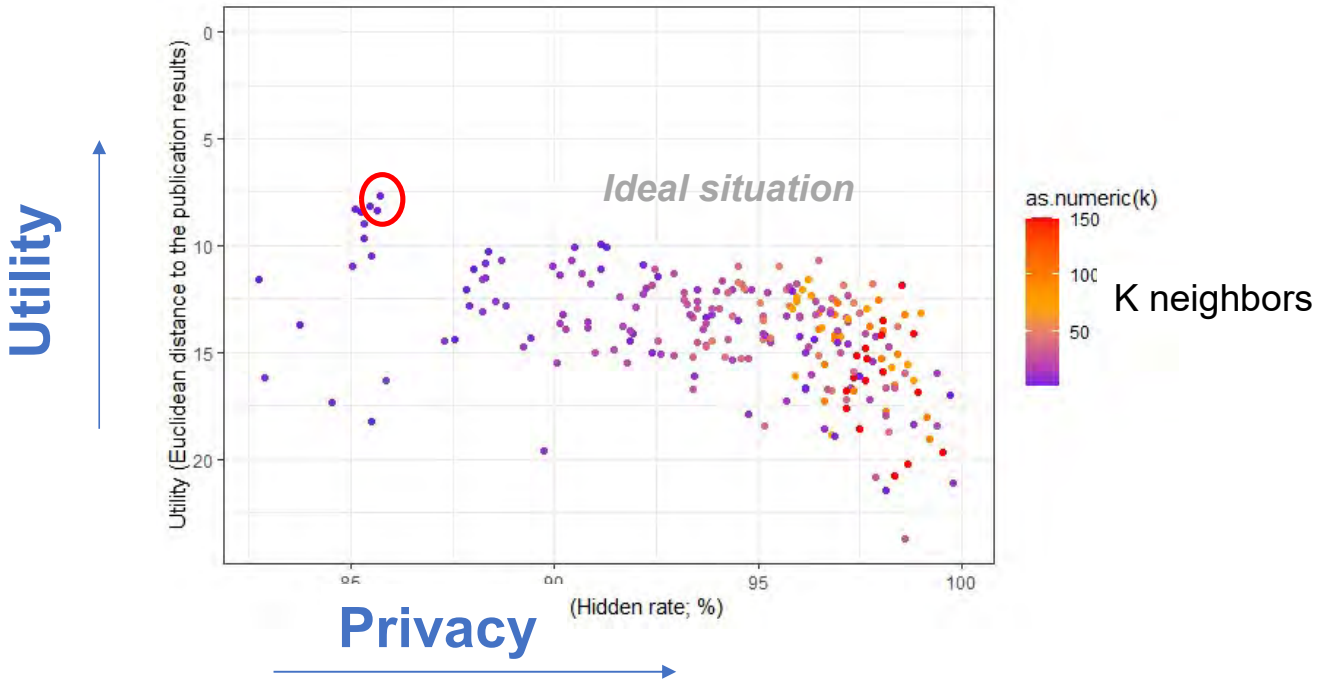
K = 5
NCP = 19 with categories reduction
HR = 82.4 %
Median LC = 2
HM = 0.04

MERCK



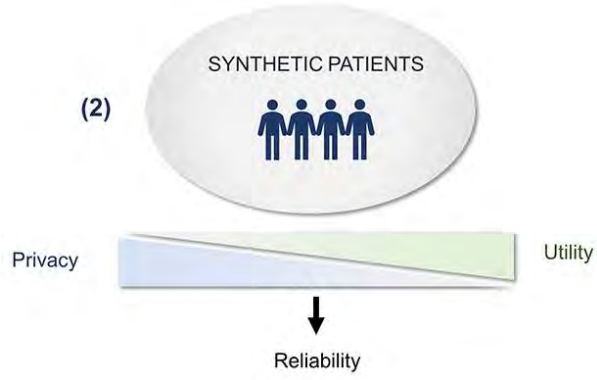
Pôle Santé UFR Médecine · DELPHI Deep Learning Proposal for Healthcare & Innovation AI & Health data · France 2030 · anr · Région Pays de la Loire · Nantes Métropole · Centre Hospitalier Universitaire de Nantes · Nantes Université · Inserm Institut national de la santé et de la recherche médicale

I) Avatarization of RCT data for multi-outcome analysis

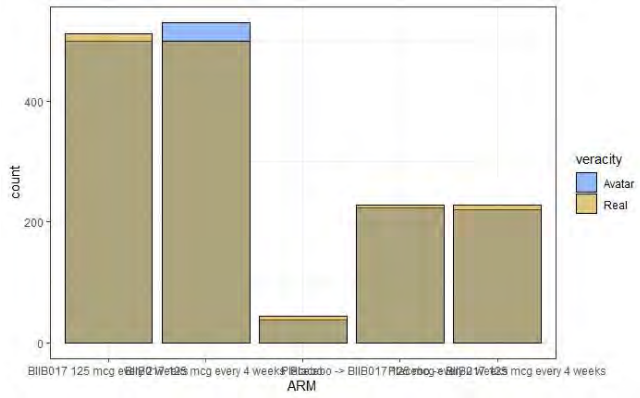


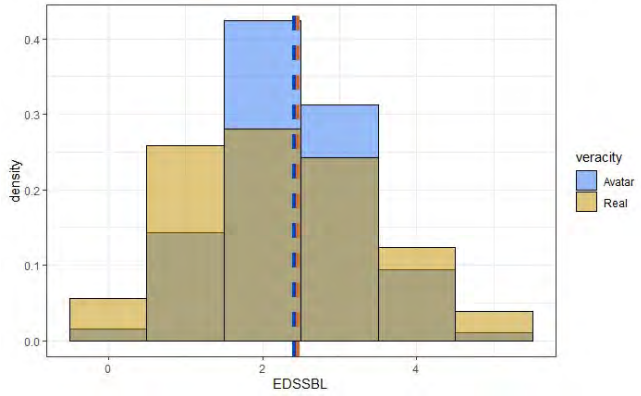
The best out Utility and Privacy

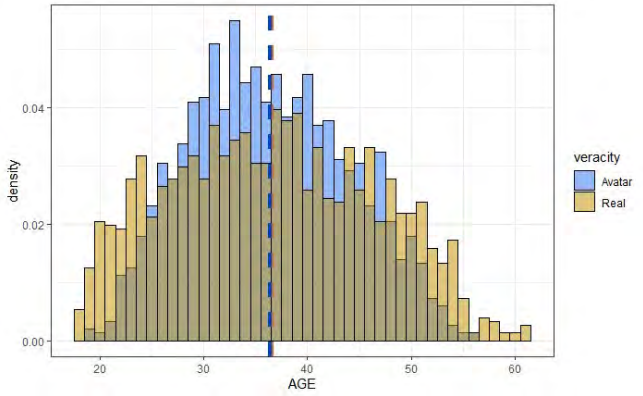
1512 observations,
25 variables (17 continuous, 8 categorical)
After dummy variables encoding: 62 columns

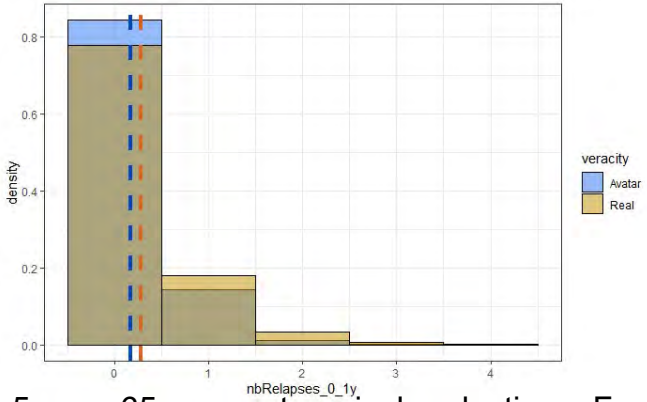


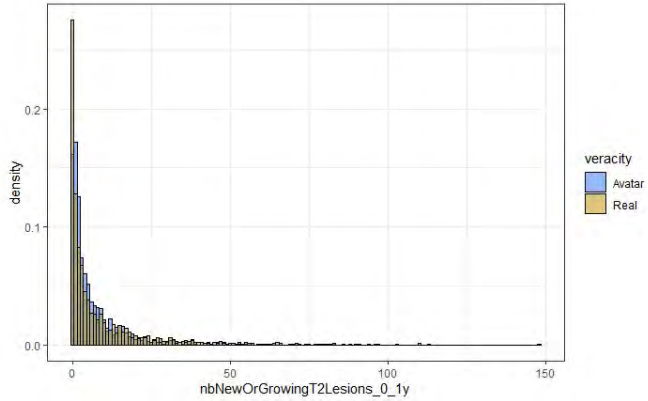
I) Avatarization of RCT data for multi-outcome analysis

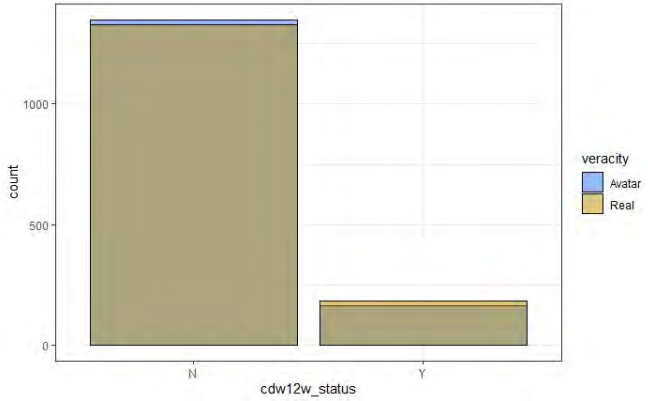












k=5, ncp=65, use_categorical_reduction = F , seed = 1

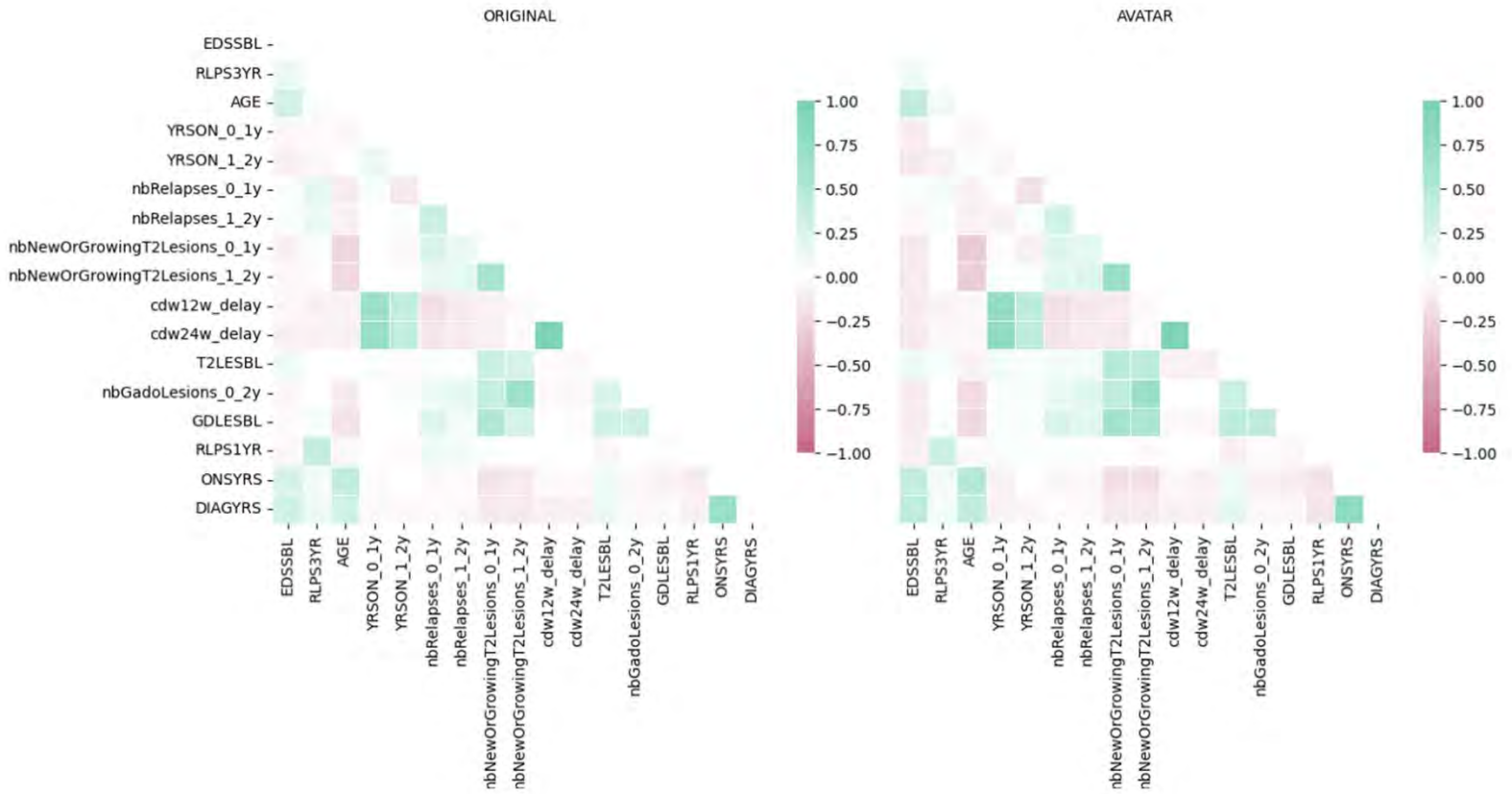
Real data Avatars

General Utility

I) Avatarization of RCT data for multi-outcome analysis

k=5, ncp=65, use_categorical_reduction = F , seed = 1

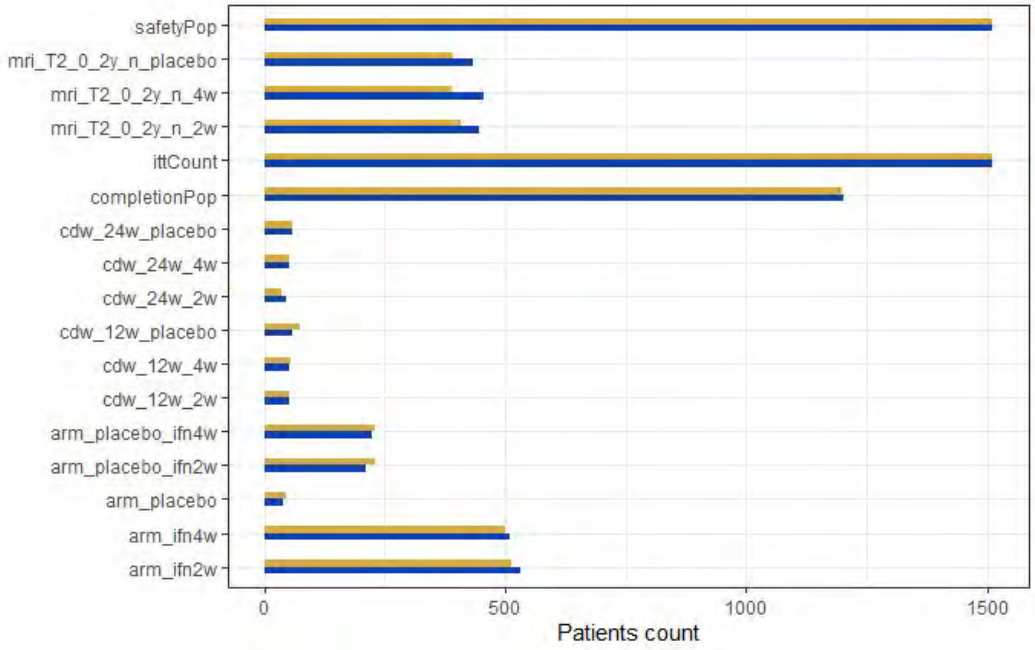
General Utility



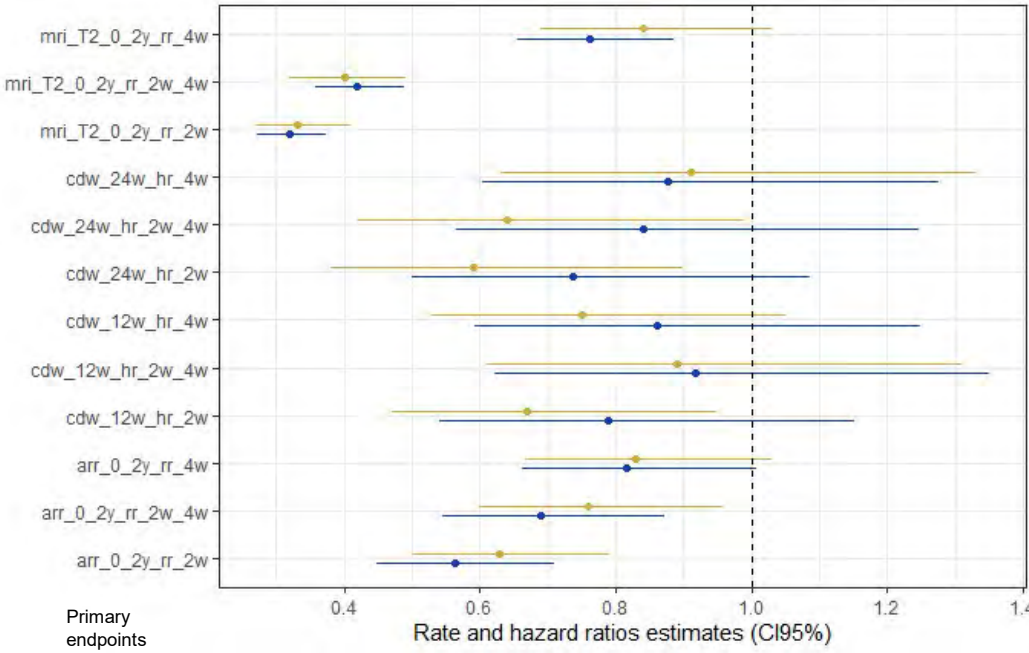
I) Avatarization of RCT data for multi-outcome analysis

k=5, ncp=65, use_categorical_reduction = F , seed = 1

Absolute counts



Association statistics and their 95%CI



General Utility

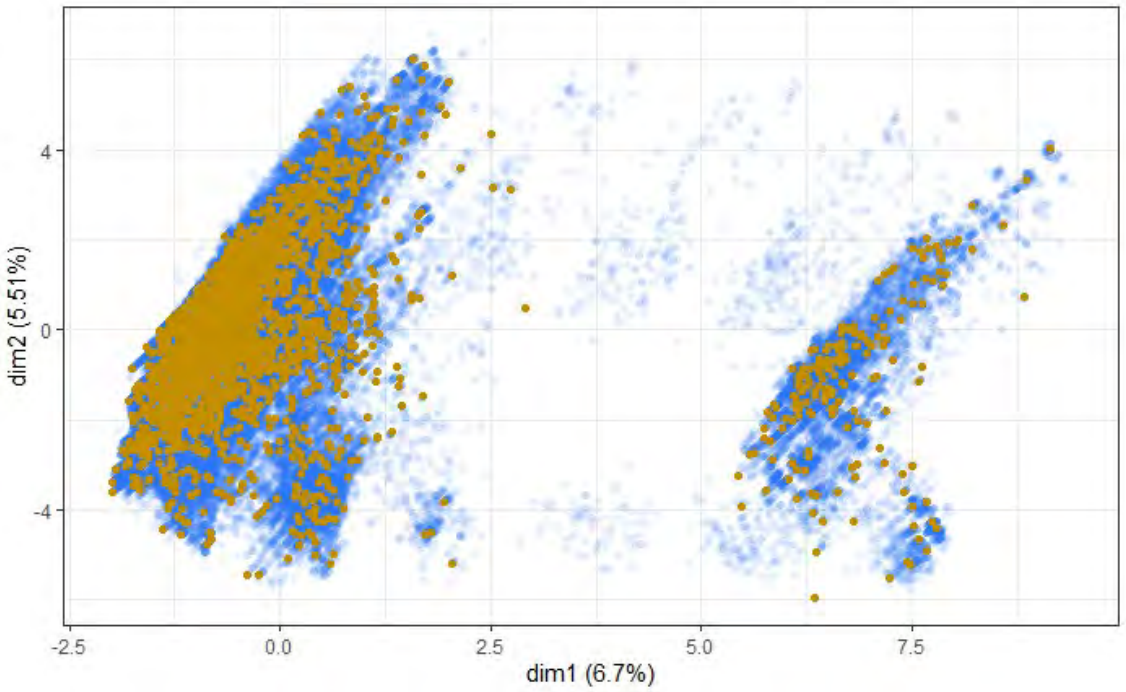
Real data Avatars

Specific Utility

II) Avatarization of OS + RCT data for filter-based similar patient data visualization (independent data points)

FAMD projection

OFSEP-HD simplified to 2744 patients



From 1-1 to 1-100 Synthetic data generation

Avatarization parameters:

K = 5
NCP = 56 (100% of inertia)
Column weights: All = 1
Use Categorical Reduction: True

Metrics:

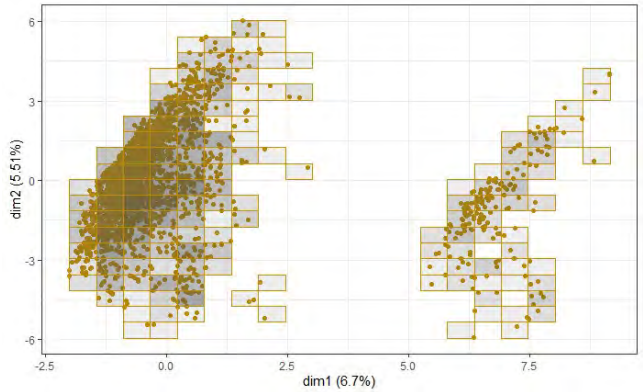
```
hiddenRate %>% summary()
medianLocalCloaking %>% summary()
closestDistanceRatio %>% summary()
hellingerMean %>% summary()
```

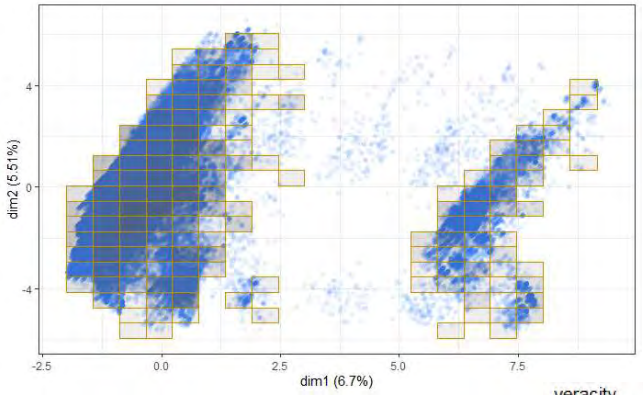
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  81.63   82.69   83.31   83.24   83.75   85.09
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      3       3       3       3       3       3
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7305  0.7436  0.7488  0.7492  0.7550  0.7679
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.03976 0.04069 0.04125 0.04132 0.04183 0.04354
```

II) Avatarization of OS + RCT data for filter-based similar patient data visualization (independent data points)

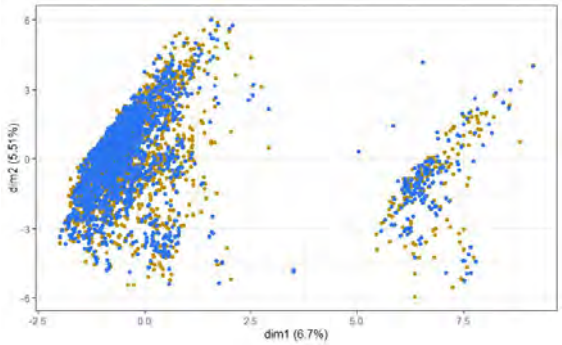
Mask in reduced multidimensionnal space

PC1 binned 20 times,
PC2 binned 20 times,
PC3 binned 20 times



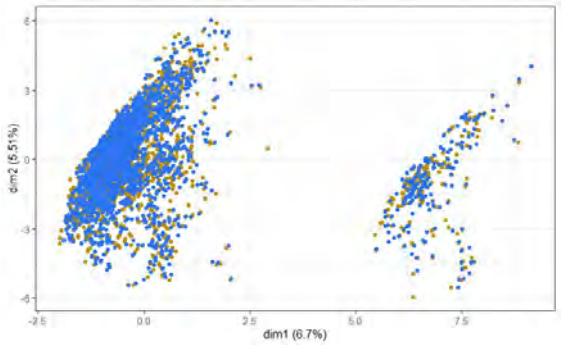


The exploratory 1:1 avatarization



→ 2733 avatars resampled against 2744 real PORs

Resampled after 100:1 avatarization



Multiple avatarizations and resampling

II) Avatarization of OS + RCT data for filter-based similar patient data visualization (independent data points)

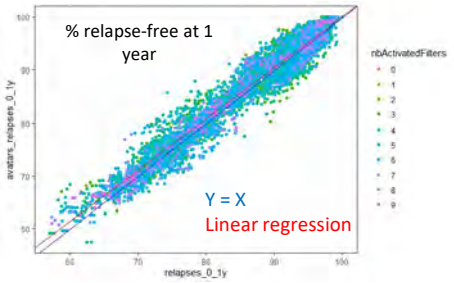
boosted

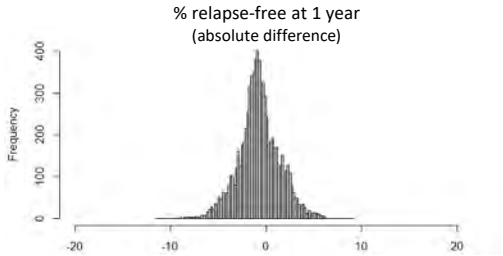
100:1 avatarization (+ mask-based resampling)

Multiple avatarizations

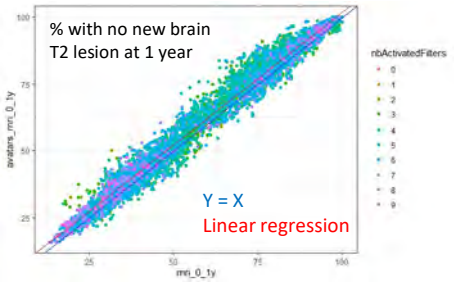
N queries = 4672
- Potential uses of CDSS
- Strata of reference patients

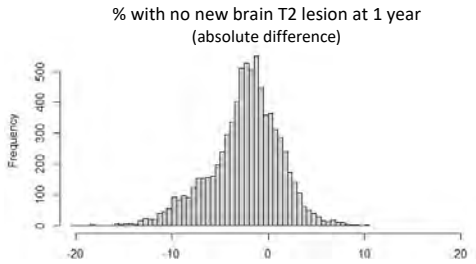
Based on all subgroups with >= 100 real patients:



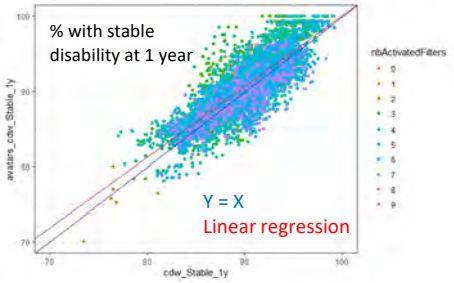


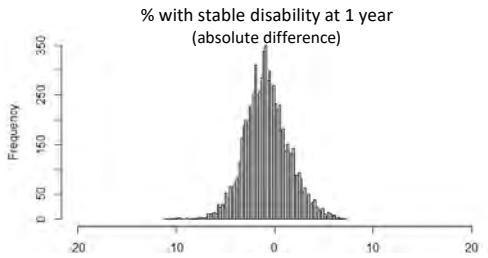
Error	Percentage of queries
Median error	1.4 %
Mean error	1.8 %
+/- 5% (absolute)	96.5 %
+/- 10% (absolute)	100 %





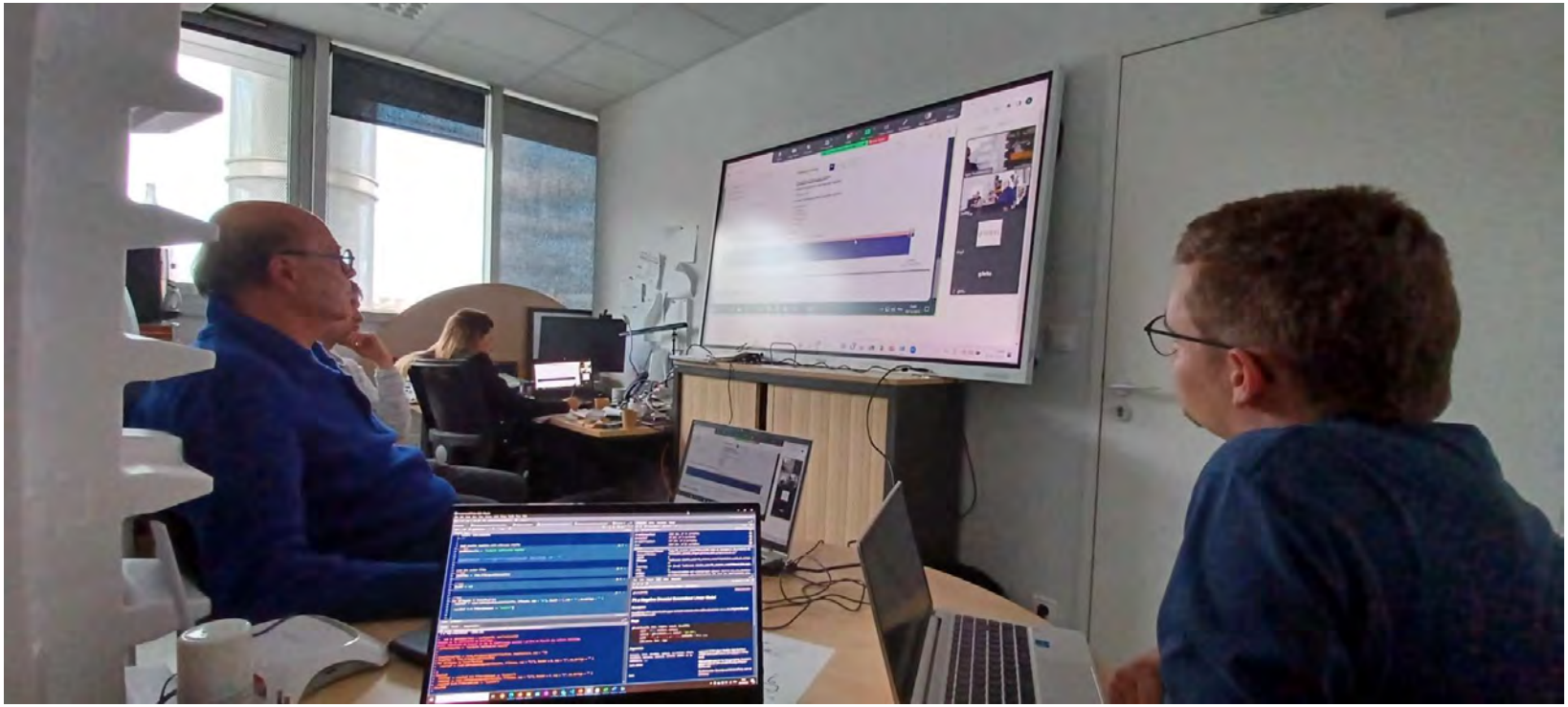
Error	Percentage of queries
Median error	2.5 %
Mean error	3.3 %
+/- 5% (absolute)	78.4 %
+/- 10% (absolute)	96.6 %





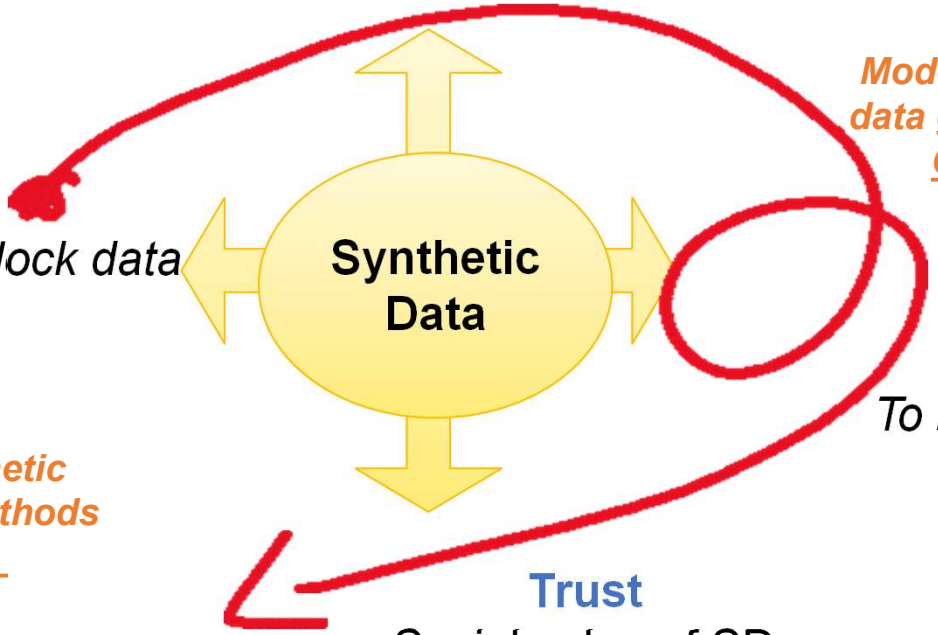
Error	Percentage of queries
Median error	1.6 %
Mean error	1.9 %
+/- 5% (absolute)	96.7 %
+/- 10% (absolute)	100 %

Tests utilisateurs et essai randomisé en cluster



Conclusion : une introspective graphique

Reality
Realistic SD
To learn via data sharing & analytics



Privacy
Anonymous SD
To respect everyone & unlock data sharing

Model-Based Synthetic data generation methods
Global models

Utility
Augmented SD
To build AI-driven algorithms

Data-Driven Synthetic data generation methods
Local Modeling

Trust
Social value of SD
To create value for all

Ouverture : Un glissement épistémologique ?

- **De la modélisation d'un phénomène … par des modèles mathématiques**
  - Ex Maladie : un malade n'est qu'une réalisation de l'entité maladie
  - Ex Jumeaux digitaux : Machine, un organe
    - L'objet de la modélisation est méta-individuel
- **De la modélisation d'un patient atteint d'une maladie à la modélisation d'un individu**
  - Recours à une modèle faible local privé à usage unique
- **Deux orientations ? Confidentialité et Utilité**
- **Deux orientations des données synthétiques**
  - F Gross "Towards a Methodology for Systems Biology" 2017)

  ***1 → Model-based** System Biology*
    - Miser sur le modèles / Les données synthétique (ex GAN)

  2 → Data-driven System Biology
    - Miser sur les donnée « data-driven »
  - Agnostique de l'analyse qui en est fait, et paramétrable (donc explicable)
  - Modélisation privée et à usage unique.

Acknowledgement:

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.

PRIMUS

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

Nantes Université

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

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

L'équipe de la Clinique des Données - janvier 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

Nantes Université

Acknowledgments

CR2TI Centre de Recherche Translationnelle en Transplantation et Immunologie UMR 1064

Team 5 « NEMO »
NEuroinflammation, Mechanisms and therapeutic Options

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

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

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

Inserm Institut national de la santé et de la recherche médicale

Nantes Université

Génération de données synthétiques centrées sur le patient, aucune raison de risquer la ré identification dans l'analyse des données biomédicales

Amphi Schwarz, IMT 11:30 - 12:00

Pr Pierre-Antoine GOURRAUD, PhD MPH

Vendredi 9 Février 24, Toulouse

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