

Annnonce des lauréats

2e appel

Programme bilatéral de recherche collaborative France-Québec L'INTELLIGENCE ARTIFICIELLE EN SANTÉ

Concours 2023-2024

28 Novembre 2023

Fonds de recherche
Santé

Québec 

Livre blanc des données synthétiques en santé

16h30 - 16h50 (CET) / 10h30 - 10h50 (EST)



Pr Pierre-Antoine Gourraud FRANCE,
CHU de Nantes – Nantes Université

Pr Michaël Chassé QUÉBEC,
CHU de Montréal – Université de Montréal



**Nantes
Université**

Université **um**
de Montréal



UQÀM

Livre blanc des données synthétiques en santé

Qui sommes-nous ?

Nantes, France

La Clinique des Données du CHU de Nantes

- Pr PA Gourraud, PU-PH
- Pr R Bourcier, Neuro-radiologiste
- Dr P Constant Dit Beaufiles, Neurologue

La DSN du CHU de Nantes

- P Guibert, Ingénieur Infrastructure

Le cluster DELPHI de Nantes Université

- J Martin-Gauthier, Cheffe de Projet

Le laboratoire CR2TI de Nantes Université

- I Faddeenkov, Ingénieur Data Scientist

Le laboratoire DCS de Nantes Université

- M Bernelin, CR Juriste CNRS

Montréal, Québec

CITADEL, Centre Hospitalier de l'Université de Mtl

- Dr Michaël Chassé, MD, PhD
- Dr Laurent Guillain Létourneau, MD, MsC, neuroradiologiste
- Dr Louis Mullie, MD et CHUM
- Katleen Blanchet, Scientifique de données
- Yue Qi, Scientifique de données

Université du Québec à Montréal (UQAM)

- Pr Sébastien Gambs, Canada Research Chair (Tier 2) in Privacy-preserving and Ethical Analysis of Big Data.

Echanges Pré-paratoires entre EDS : Ex Cancer Prostate (C61)

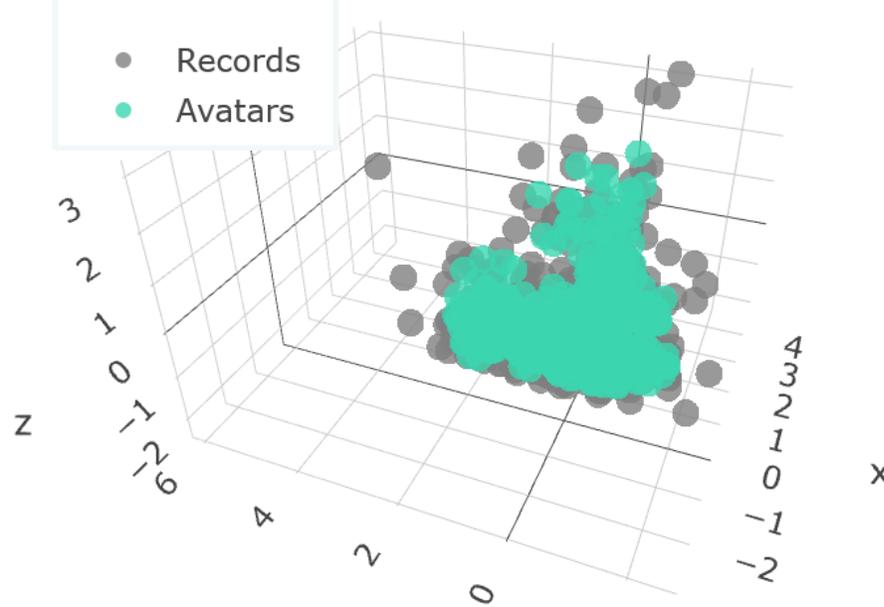


413 patients



Guillaudeau et al 2023 Nature Digital Medicine

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



npj | Digital Medicine



413 données synthétiques anonymes
"avatars de patients"

Sexe	Age_visite_1	Date_visite_1	Nb_visites
M	72	13/08/2022	6
M	68	15/04/2022	10
M	70	25/10/2022	3
M	70	12/02/2022	2
M	82	18/06/2022	3
M	66	01/04/2022	9
M	81	23/04/2022	5
M	68	31/01/2022	13
M	71	08/02/2022	2
M	80	29/07/2022	6
M	70	03/03/2022	14
M	88	26/05/2022	5
M	79	02/02/2022	10
M	70	12/03/2022	19
M	61	31/01/2022	7
M	66	22/03/2022	4
M	86	20/10/2022	1
M	59	30/07/2022	5
...
M	68	07/01/2022	6



Summary



Privacy metric	Value	Target
Hidden rate	90.8 %	> 90 %
Median local cloaking	4.0	> 5
Distance to closest	0.2	> 0.2
Closest distance ratio	0.77	> 0.3
Categorical hidden rate	100.0 %	> 90 %

Sec-exp 1-4

Librairie Clover

🏠 Clover

Search docs

- generators
- metrics
- utils
- tests
- notebooks

🏠 / Welcome to Clover's documentation!

[View page source](#)

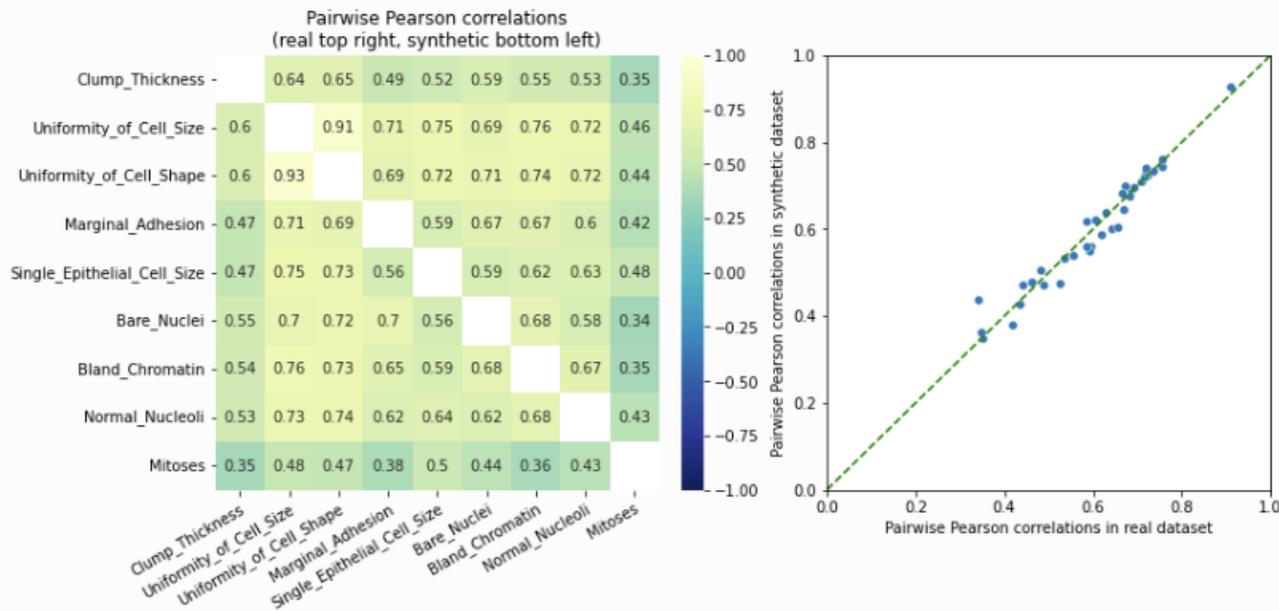
Welcome to Clover's documentation!

Synthetic Health Data Generation and Validation Library

Advances in health research are constrained by the availability of data. Indeed, access to a large amount of data from different sources is a key factor to increase the generalizability of the machine learning algorithms and validate them and thus improve healthcare for the population.



Metric: Pairwise Correlation Difference



]:

	name	objective	min	max	submetric	value
Categorical Consistency		max	0	1.0	0	within_ratio 1.000000
	Categorical Statistics	max	0	1.0	0	support_coverage 1.000000
Classification	1					frequency_coverage 0.975110
		min	0	1.0	0	diff_real_train 0.003894
	1					diff_synth_train 0.002164
Continuous Consistency	2					diff_real_synth 0.003207
		max	0	1.0	0	within_ratio 1.000000
Continuous Statistics		min	0	inf	0	median_l1_distance 0.012346
	1					iqr_l1_distance 0.024691
Cross Classification		min	0	1.0	0	real_train 0.003147
	1					synth_train 0.000779
Cross Regression		min	0	inf	0	real_train 0.541089
	1					synth_train 0.158309
Distinguishability		min	0	1.0	0	propensity_mse 0.037861
	1					prediction_mse_real 0.066487
	2					prediction_mse_synth 0.058752
FScore	3					prediction_auc 0.000000
		min	0	inf	0	diff_f_score 0.114737
Feature Importance		min	0	inf	0	diff_permutation_importance 0.019435
Categorical Univariate Distance		min	0	1.0	0	hellinger_distance 0.018309
Continuous Univariate Distance		min	0	1.0	0	hellinger_distance 0.043674
Categorical Univariate Distance		min	0	inf	0	kl_divergence 0.001348
Continuous Univariate Distance		min	0	inf	0	kl_divergence 0.008203
Pairwise Correlation Difference		min	0	inf	0	norm 0.245490

Livre blanc des données synthétiques en santé

Qu'est ce que le projet et pourquoi ?



Conception d'un guide de bonnes pratiques destiné aux **responsables de traitement** (fiduciaires de données) pour mieux accompagner la génération de **données synthétiques anonymes en santé**, en se concentrant sur leur **sécurité, informativité** et **authenticité** avec un cas d'usage pour la **caractérisation des facteurs d'anévrisme intracrânien**.

<https://gitlab.com/healthdatahub/tutoriel-generation-de-donnees-synthetiques-en-sante>

Objectif 1.2: Feuille de Route du HDH Produire des données anonymes ou synthétiques associées aux bases du catalogue HDH



- 1. Convaincre** de l'intérêt de cette forme d'IA par des **métriques** de validations et des protocoles d'attaques d'inférence.
- 2. Améliorer** la caractérisation de l'**anévrisme intracrânien** en France et au Québec.
- 3. Rapporter** les différences entre les **cadres juridiques** canadiens, européens et français en matière de protection de la vie privée, de gouvernance des données, et de normes de sécurité.

Accélérer la recherche collaborative internationale pour améliorer les soins de santé sans compromis avec la vie privée.

Livre blanc des données synthétiques en santé

Méthodologie

G29
N°05/
2014



Génération de données synthétiques
8 méthodes



Évaluation de la qualité des données et des générateurs
Utilité / Confidentialité



Caractérisation des ruptures d'anévrismes
QC ↔ FR

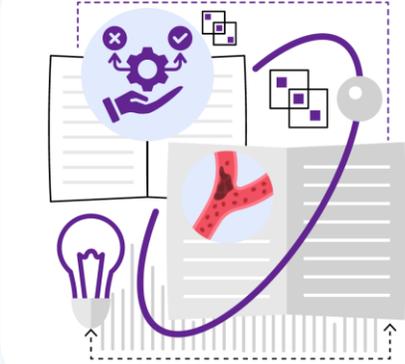


Livre blanc des données synthétiques
ELSI

Janvier 2024 – Juin 2025

Livre blanc des données synthétiques en santé

Résultats



Publications

- Meilleure caractérisation des facteurs de rupture des anévrismes
- Librairie de référence en **accès libre** de génération et validation (utilité et vie privée) des données synthétiques et générateurs (codes et notebooks publiés sur GitHub et le HDH)
- Livre blanc décrivant les bonnes pratiques pour générer des données synthétiques interopérables en santé impliquant un volet réglementaire



Formations

- Quatre étudiants de niveau master (2 data scientists, 1 juriste, 1 interne en médecine) à la génération de données synthétiques.
- Événement de dissémination bilingue à rayonnement international.

Livre blanc des données synthétiques en santé

Des questions ?

DELPHI@univ-nantes.fr