Nantes Université

L'innovation autour de la donnée (massive)

(1) La Clinique des donnée et le réseau de CDC; (2) Objet connecté; (3) Données synthétiques

Prof. Pierre-Antoine Gourraud, Nantes Université & CHU

15 Juin 2022, Faculté de Pharmacie

HUGO
Hopitaux Universitaires Grand Ouest

- **COI :**

PA Gourraud is the founder of Methodomics (2008) and the co-founder of Big data Santé (2018). He consults for major pharmaceutical companies, all of which are handled through academic pipelines (AstraZeneca, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Methodomics, Merck, Mérieux, Sanofi-Genzyme, Octopize). PA Gourraud is volunteer board member at AXA mutual insurance company (2021). He has no prescription activity with either drugs or devices.

Nantes Université

"Our generation will be called naïve about data"

- ***"Nous sommes des naïfs de la donnée"***
- ***PA Gourraud Y Coatanlem***
- ***Le Monde Oct 5th 2021***

« Avec des protocoles d'accès plus souples, les données publiques pourront constituer un gisement de valeur du XXIe siècle »

Tribune 05.10.2021

Le Monde

Yann Coatanlem

Président du club de réflexion Praxis

Pierre-Antoine Gourraud

Professeur à la faculté de médecine de l'université de Nantes

Les moyens existent de libérer l'exploitation des données tout en protégeant la confidentialité, notamment pour le croisement et le partage des fichiers de vaccinations et de tests, expliquent, dans une tribune au « Monde », Yann Coatanlem, président du club de réflexion Praxis et Pierre-Antoine Gourraud, professeur de médecine.

Publié aujourd'hui à 06h15 Temps de Lecture 4 min.

Tribune. Nous sommes des naïfs de la donnée ! Bien que plus ou moins conscients que les données sont au XXIe siècle l'équivalent de la terre arable à l'ère agricole ou de la machine au XIXe siècle, nous n'exploitons encore qu'insuffisamment les gisements d'opportunités dans ce domaine. Aujourd'hui, en pleine crise de Covid-19, le croisement et le partage des fichiers de vaccinations et de tests posent encore problème alors même que les enjeux de santé publique sont criants. C'est donc un véritable aggiornamento des politiques en la matière que nous appelons de nos vœux.

Dans les débats publics, les enjeux sont malheureusement souvent confondus : enjeux de confidentialité, d'usage (la finalité de l'analyse des données), d'usages secondaires (par opposition à l'intentionnalité première des données), de contrôle des usages (quelles données, pour faire quoi), de contrôle des usagers (par qui), de sensibilité (quelles sont les conséquences potentielles de l'interprétation des données). Cette confusion nuit à la transparence, à la collecte, à l'organisation, à la valorisation des données. Elle nuit finalement à la confiance requise pour que le développement économique se nourrisse de la création et de la diffusion des connaissances.

Nantes Université

Introduction to the true nature of (medical) Data

euni well

The "Ode to Joy"

- Composed as a choral symphony
  - …. By L Beethoven
- Played
  - .. . By an symphonic orchestra
- Written as a poem
  - … By Friedrich von Schiller
- Officially interpreted
  - .. By Herbert von Karajan
- Translated in French English
  - … By many
- Sung
  - … by anyone who dares
- Et caetera …

Electronic Medical Records

- Acquired
  - …. From patients
- Written
  - .. . By caregivers
- Produced
  - … by medical devices
- Paid
  - … by Insurance companies
- Stored
  - … by Care Institutions
- Transformed
  - … by data scientists
- Et caetera …

Personal Medical Data is not similar to material good
We need to take "good care" of it…

Pôle Santé
UFR Médecine

Nantes Université

Enjeux de données de santé : Définitions : Données Massives

Pôle Santé UFR Médecine

> Définition par source :

Données d'hospitalisation / consultation	Consommation de soins	Objets connectés	Produites à des fins de recherche	Tout autre type de données
• Diagnostic • Traitement • Biologie • Imagerie • Génétique • …	• Prescriptions, • Données de l'Assurance Maladie • …	• Médicaux : tensiomètre, implant cardiaque, … •Non médicaux : montre connectée, pèse-personne…	• Omics : génomique, transcriptomique… • Microbiote • Cohortes •Biocollections •…	• Données socio-démographiques • Géographiques • Météorologiques • …

> Définition par structure :

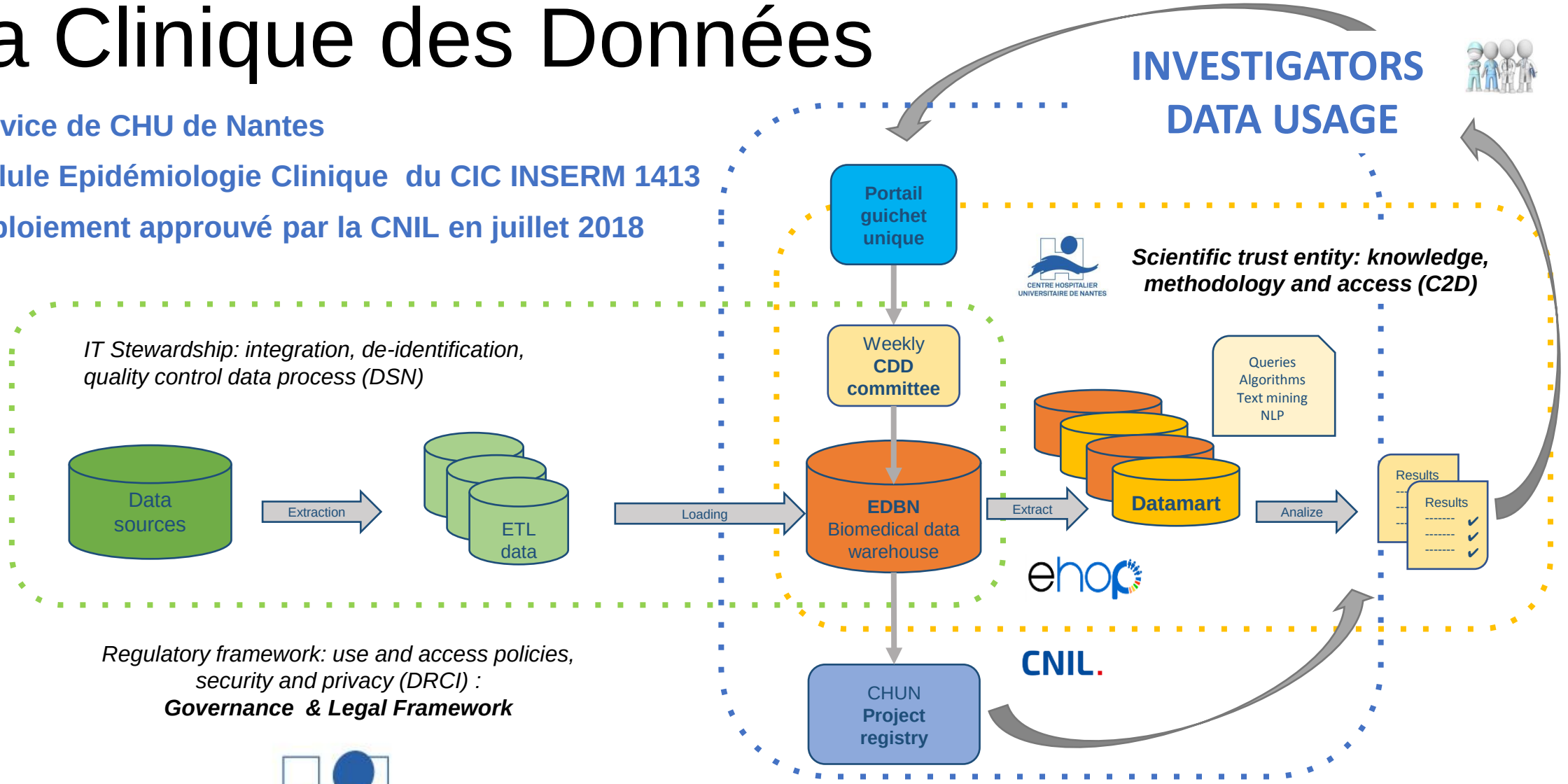
Données structurées	Données non structurées
• Valeurs biologiques (hémoglobine, glycémie…) • Données génétiques • PMSI • …	• Compte-rendu : Texte « brut » -> TALN • Données de signal (ECG) • Images (scanner…) • …

Partie 1. « Centre des données clinique »
l'accessibilité des données issues du soin au sein de HUGO

Screening ; « données de vie réelles » , « recyclage de données issues du soin » --- > Pas d'abord une problème de technique

La Clinique des Données

- **Service de CHU de Nantes**
- **Cellule Epidémiologie Clinique du CIC INSERM 1413**
- **Déploiement approuvé par la CNIL en juillet 2018**



Périmètre plus large que les entrepôts de données issues du soin ré-exploités à des fins de recherche

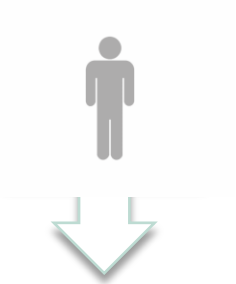
		RESPONSABILITÉ LÉGALE *de la source de données : portée par le CHU de Nantes*	
		OUI	**NON**
INTENTIONNALITÉ *Données collectées à visée de soin*	**NON**	Entrepôt de données de santé Données médico-administratives et de soins 	 **+ Projets « exceptionnels » :** Météo France, SOS médecins, Argos…
	OUI	**Cohortes du CHU de Nantes :** BRUGADA, ICAN, EASY, VISIOCORT, EXAN, COVER, VALIDate, CoHPT, IT-DIAB, CORONADO…	**Cohortes Nationales ou Internationales Bases de données externes**  

La Clinique des Données *« Centre des données clinique »*

- **Service de CHU de Nantes**
- **Cellule Epidémiologie Clinique du CIC INSERM 1413**
- **Déploiement approuvé par la CNIL en juillet 2018**
  - **Dont Entrepôts de données Biomédicales issues du soin**
  - **407 projets pris en charge.**

Governance Matrix		*Legal Resp. (CHUN)*	
		Yes	**No**
Research Intentionality	**Yes**	Registry Cohorts	Nat. or Int. Databases
	No	**BDW**	SNDS

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES



1,5 millions Documented Patients

540 millions Structured Data

34 millions Textual Documents

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale

Nantes Université

Les données : c'est le vecteur de transformation du XIXème siècle

« A terme, il n'y aura plus un projet de recherche en santé qui ne pourrait pas bénéficier d'une extraction de données d'entrepôt hospitalier. »

Nantes Université

Illustrations

Projet SEVASAR

INTENTIONNALITÉ	RESPONSABILITÉ OUI	RESPONSABILITÉ NON
OUI		
NON	X	

Identification des patients atteints du variant anglais du COVID-19

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

- Projet national, problématique d'identification des patients avec le variant anglais hors CHU

Support CdD

- **Screening :** Identification population cible grâce à la recherche textuelle sur l'EDS, mise à disposition d'un listing d'IPP

Nantes Université

Illustrations

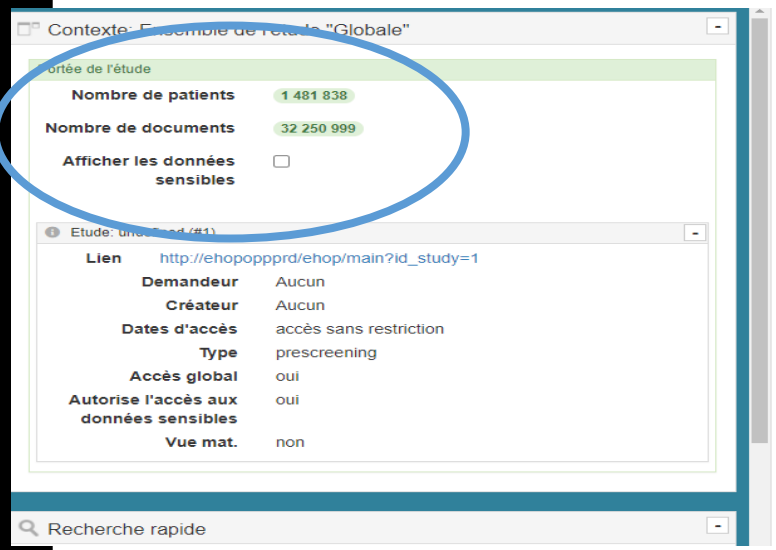
Projet SEVASAR

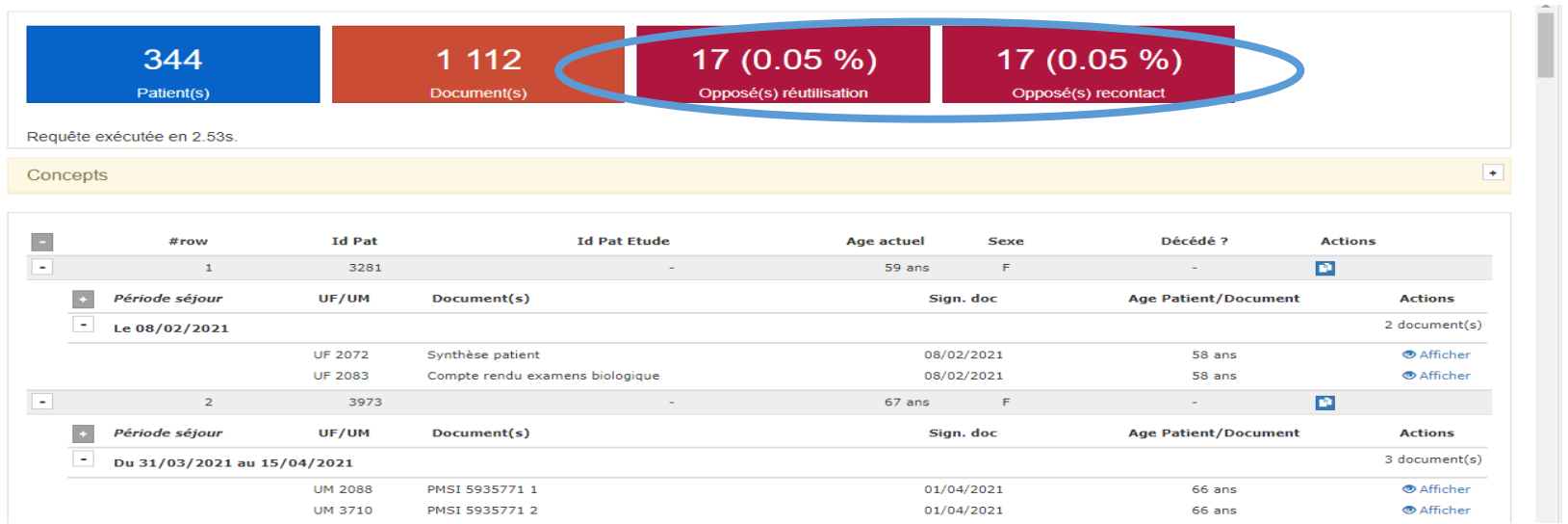
Identification des patients atteints du variant anglais du COVID-19

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

	RESPONSABILITÉ OUI	RESPONSABILITÉ NON
INTENTIONNALITÉ OUI		
INTENTIONNALITÉ NON	X	

3. Recherche eHOP par interface graphique : résultats





Nantes Université

Illustrations

Projet SEVASAR

Identification des patients atteints du variant anglais du COVID-19

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

INTENTIONNALITÉ \ RESPONSABILITÉ	OUI	NON
OUI		
NON	X	

Recherche eHOP par interface graphique : CR également accessible

Nature: Ecouvillon nasopharyngé

RECHERCHE DE VIRUS PAR BIOLOGIE MOLECULAIRE

Recherche du SARS CoV-2 (COVID-19):

Test moléculaire rapide IDNOW COVID-19 (ABBOTT)

Résultat: POSITIF

Présence d'ARN viral compatible avec une excrétion virale significative; patient à considérer comme contagieux.

SARS CoV2: recherche de la mutation E484K

Kit VirSNiP SARS CoV-2 spike E484K (TIB MOLBIOL)

Mutation E484K: NON

SARS CoV2: recherche de la mutation N501Y

Kit VirSNiP SARS CoV-2 spike N501Y (TIB MOLBIOL)

Mutation N501Y: OUI

Interprétation: Détection d'un variant dit "anglais"

CONCLUSION: Mise en évidence d'un coronavirus SARS CoV-2 (COVID-19)

Nantes Université

Illustrations

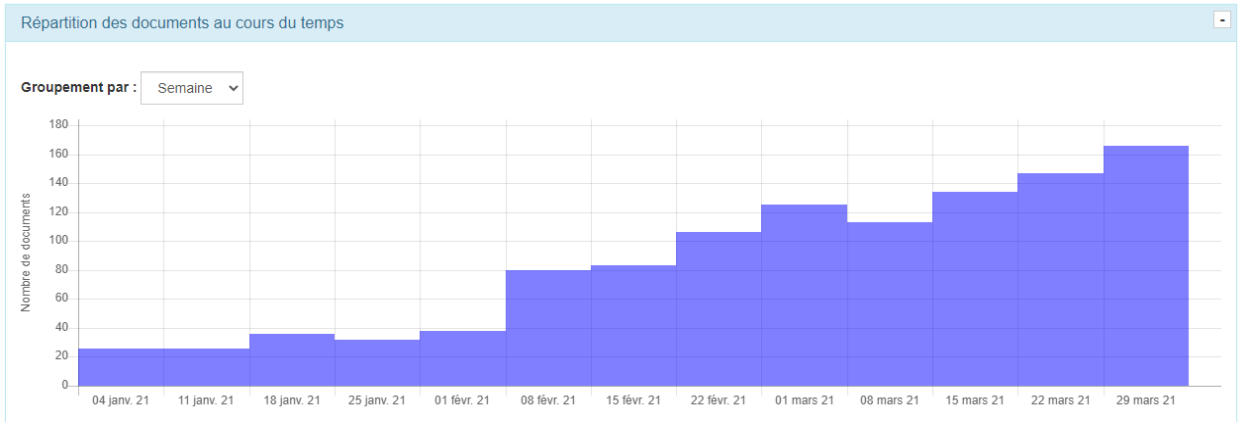
Projet SEVASAR

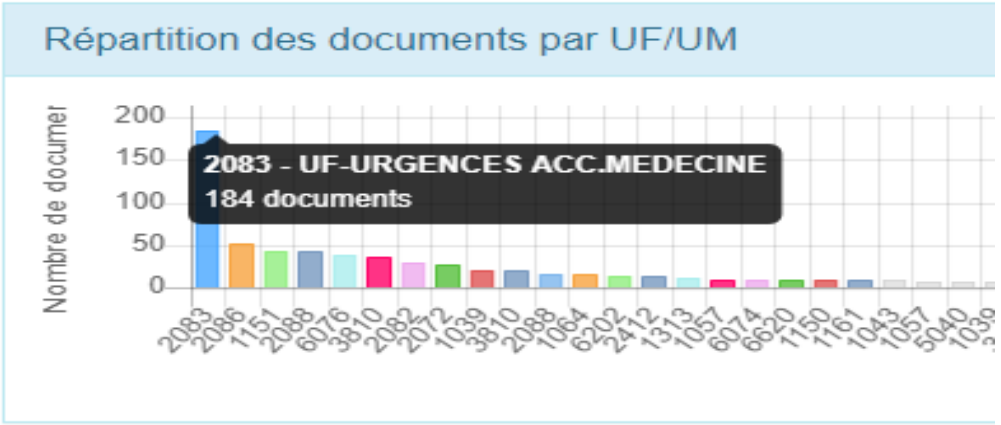
Identification des patients atteints du variant anglais du COVID-19

RESPONSABILITÉ	OUI	NON
INTENTIONNALITÉ OUI		
INTENTIONNALITÉ NON	X	

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

Recherche eHOP par interface graphique : Distribution temps et UF/UM





Nantes Université in West of France

2016 : From Local Biomedical Data Warehouses

- **Development et optimization** :

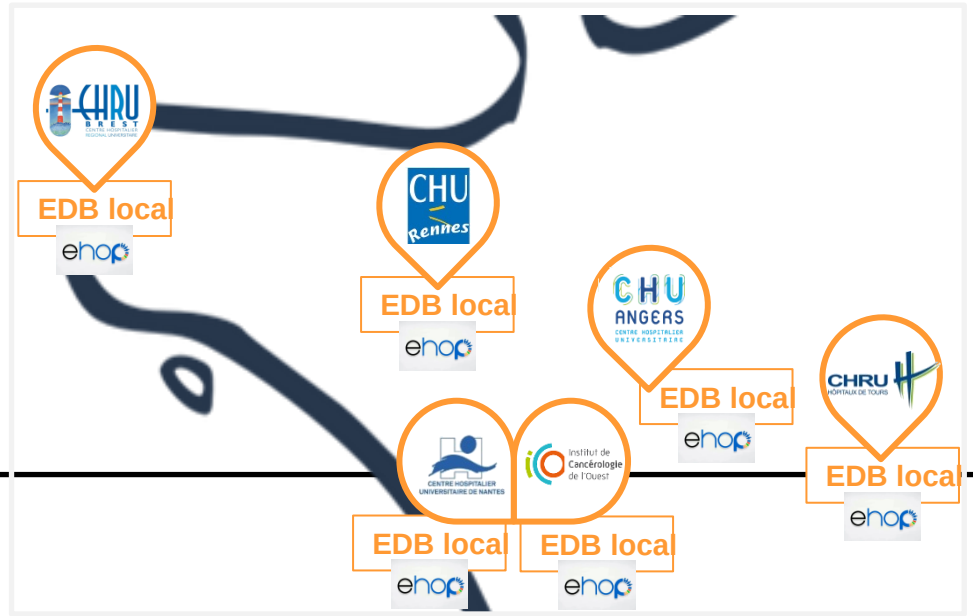
  1- Data Governance

  2- Legal framework

  3- Technical Issue.

Pr Marc CUGGIA

Technical Deployment - public- private partnership

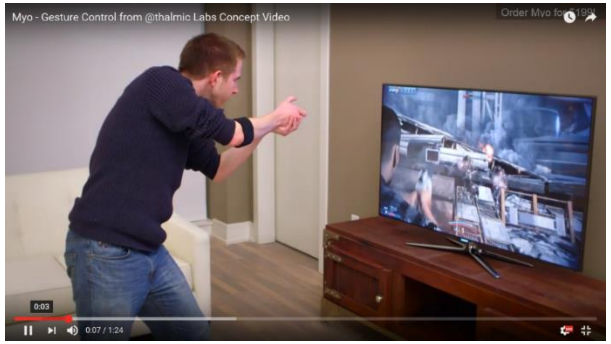


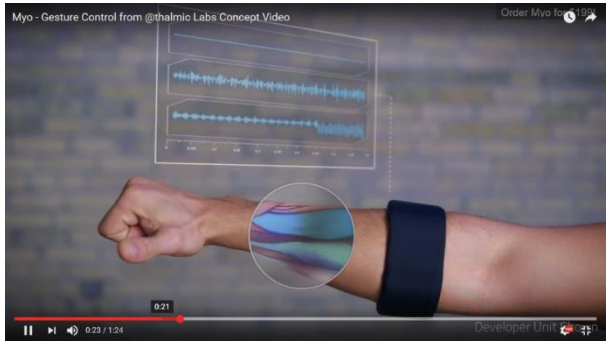
Partie 2. Exemple de recherche clinique avec un objet connecté…

Effondrement du coût de collecte de la data & un projet disruptif ?

Electro myographe de Surface au service de la collecte de données

- Examen Neurologique traditionnel
  - Observation : pratique de l'art de la médecine
  - Test de Coordination : système moteur & sensoriel
    - Lesions cérébellesue et/ou vestibulaire
- **Présent :** Digitalisation des pratiques





Electro-Myographe de Surface au service de la collecte de données

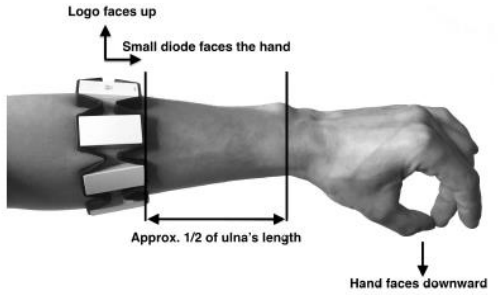
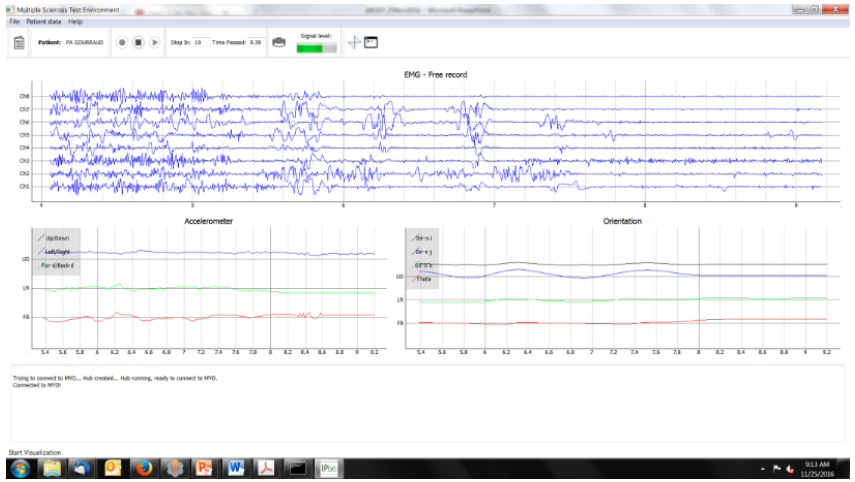


Fig. 2. Armband positioning on the forearm, example shown for finger tapping test.



- **Nouvelles données nouveaux Outils :**
  - Algorithme d'intelligence artificielle"
  - Traitements des données à grande échelle
  - Classement des patients
- **Modification de notre regard sur la sclérose en plaques**
  - Analyse de la marche

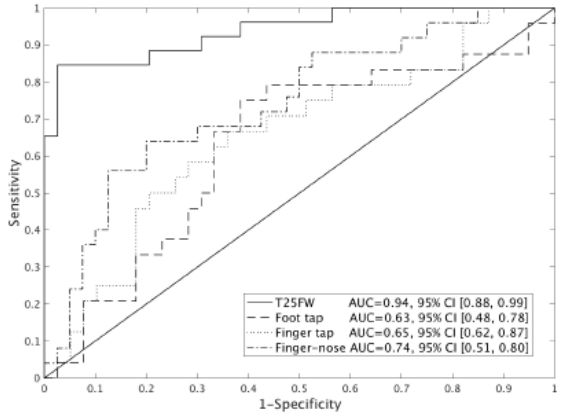


Fig. 4. ROC curves for four motor function tests, T25FW stands for timed 25 foot walk.

SVM-based Tool to Detect Patients with Multiple Sclerosis Using a Commercial EMG Sensor

Konstantin Akhmadeev*, Aya Houssein†, Said Moussaoui†, Einar A. Høgestøl‡, Ingerid Tuttuen‡, Hanne F. Harbo‡, Steffan D. Bos-Haugen‡, Jennifer Graves¶, David-Axel Laplaud‖, Pierre-Antoine Gourraud‖

*Université de Nantes, LS2N UMR 6004, Nantes, Frances, Email: konstantin.akhmadeev@ls2n.fr
†Ecole Centrale Nantes, LS2N UMR 6004, Nantes, France
‡Department of Neurology, Institute of Clinical Medicine, University of Oslo, Oslo, Norway
¶Department of Neurology, School of Medicine, University of California San Francisco, San Francisco, CA, USA
‖Université de Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie UMR 1064, Nantes, France

Abstract—Multiple sclerosis (MS) is a major auto-immune disease that is the leading cause of non-traumatic impairment of the central nervous system (CNS) in young adults. Successful treatment of MS patients depends on accurate tools for both the MS diagnosis and the disability progression. In current and upcoming studies the authors aim to explore the capabilities of applying a commercial electromyographic and inertial sensor (MYO Armband by Thalmic Labs Inc.), coupled with a multichannel signal processing tool, to standard neurological examination. In this pilot study we formulate a two-class "healthy control" - "having MS" classification problem. A dataset of electromyographic signals and inertial sensor measurements from 71 individuals (31 MS patients and 40 healthy controls) was acquired during standard neurological examination routine. Temporal and spectral features of the signals were extracted in order to train and validate a classification model. Finally, a Support Vector Machine classifier was obtained giving AUROC = 0.94, 95% CI = [0.88, 0.99] and verified using 8-fold cross-validation. We propose a set of signal descriptors that correlate with objective components of the neurological examination. The proposed signal acquisition and processing technique, being easy to integrate into the traditional neurological exam, may have high potential for aiding in diagnosing MS and quantifying its progression.

I. Introduction

Multiple sclerosis (MS) is a chronic debilitating neurological disorder that mainly affects young individuals aged between 20 and 40. As a cause of neurologic disability MS is second only to trauma, having its prevalence estimated at 2.5 million worldwide in 2014. The actual cause of MS is yet to be identified, but a complex interaction between genetic and environmental factors contributes to the risk. To date, there is no reliable method to predict MS onset or progression. Successful managing of the symptoms and attacks for MS patients highly depends on an accurate and timely diagnosis as well as the possibility to measure disability progression.

Diagnostic criteria for multiple sclerosis include a number of clinical and paraclinical laboratory assessments [1], [2]: cerebrospinal fluid analysis, study of visual evoked potentials, electromyography analysis, neuroimaging and motor/sensory/balance function tests. The latter involves various motor tasks to be accomplished by the subject: timed 25-foot walk [3], 9-hole peg test, finger-to-nose test [4], heel-knee-shin test, finger tapping, foot tapping, etc.

The most common motor manifestations of MS are muscle fatigue, spasticity and tremor. Listed symptoms involve abnormal functioning of skeletal muscles and thus affect their activation patterns. In such cases, deviations may be revealed by analysis of limb trajectories and of involved muscles' electromyography (EMG). These measurements are proven to be efficient in different studies of MS progression [5]–[7]. Thus, an EMG recording along with the inertial measurement unit (IMU) data may aid to characterise presence and severity of MS.

Common MS diagnosis and progression study approaches, as those listed above, require specific equipment, procedures and clinical expertise. A lack of them may slow down or make the diagnosis impossible, which is a common case for low populated areas or developing countries. A possible way to overcome these difficulties is to apply a widespread cheap acquisition system, along with unified assessment protocol and automated decision-making. As such an acquisition system we propose the MYO armband (figure 1) developed and commercialised by Thalmic Labs Inc [8]. It comprises eight EMG channels and an IMU giving acceleration, orientation and rotation speed measurements in three axes. This device is wireless, cheap, easy to use, actively supported by community and can be shipped to any location.

MYO armband's default software is capable of recognising five different hand gestures, based on EMG. Also, IMU sensor provides a pointer control. In academic studies, this device was applied in sign language gesture recognition [9] and prosthetic control [10], [11]. Typical signal processing pipeline in these applications consists of the following steps: windowing, feature extraction, dimensionality reduction and classification using machine learning techniques [12], [13]. Such an approach may also be effective in an application to MS diagnostics since there is no strictly defined model of how MS affects surface EMG signals or limb trajectories. Other reasons to use machine learning techniques in this case are the dimensionality of the data and the fact that measurements

Nantes Université

Quelles leçons dans la Santé?

1. **Au centre du développement du numérique la question des usages**
   - Pas « disruptif » - Pas une révolution.. Des courbes de croissance version 2.0
   - Que faire de ces « données »? Est-ce qu'on en a trop?
   - Prolonge l'observation et la réflexion ne la remplace pas
2. **Pas de remplacement du soignant mais Modification de la relation soigné soignant**
   - Intrusion d'un tiers pole
   - Rôle croissant des données
   - "À la demande" – de manière "personnalisée"
     - Point de départ du calcul est un individu donné
3. **Double changement dont nous ne mesurons pas encore la profondeur des conséquences**
   - 1 La connexion aux bases de données
   - 2 La capacité à calculer à la demande

Partie 3. Des données à réinterpréter…
Des données à transformer pour leur donner de la valeur et les libérer

Comme une partition de musique … un exemple ultime faire sortir les données du RGPD : passer de la donnée pseudonymisée potentiellement ré-identifiant à de la données synthétiques.

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Inserm
Institut national de la santé et de la recherche médicale

Nantes Université

La confiance des patients

«Quand on analyse des données , il n'y a plus de justification à faire courir un risque de ré-identification aux patients. »

Nantes Université

Actualité Légale

« Naïf de la donnée »
Rôle d'alerte et d'éveil de la CNIL

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

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**.

Nantes Université

Using synthetic Data in biomedical data warehouse

OCTOPIZE
MIMETHIK DATA

- **User point of view.**
  - On premise deployment
  - Nantes university Hospital : Biomedical Data warehouse
- **Publication Rousseau et al. 2020**
  - Open Source algorithm …
    - https://github.com/ICAN-aneurysms/RIA-predict
  - Open Data in biomedical research …
    - But … Risk of re identification
      - » Article 29 working party
    - A shift from Simulated data …
    - to Synthetic data; avatars
- Focus 1 – Re-identification metrics
- Focus 2 - Statistical value of synthetic data

Cerebrovascular disease

ORIGINAL RESEARCH

Location of intracranial aneurysms is the main factor associated with rupture in the ICAN population

Olivia Rousseau,[1] Matilde Karakachoff,[1] Alban Gaignard,[2] Lise Bellanger,[3] Philippe Bijlenga,[4] Pacôme Constant Dit Beaufils,[1] Vincent L'Allinec,[5] Olivier Levrier,[6] Pierre Aguettaz,[7] Jean-Philippe Desilles,[8] Caterina Micheloz[9], Gaultier Marnat,[10] Anne-Clémence Vion,[2] Gervaise Loirand,[2] Hubert Desal,[11] Richard Redon,[2] Pierre-Antoine Gourraud,[1] Romain Bourcier[11] The ICAN Investigators

For numbered affiliations see end of article.

Correspondence to
Dr Romain Bourcier, Department of Diagnostic and Therapeutic Neuroradiology, University Hospital of Nantes, L'Institut du thorax, Inserm 1087, CNRS, UMV Nantes, University Hospital Centre Nantes, Nantes, Pays de la Loire, France; romain.bourcier2@gmail.com

Received 24 June 2020
Revised 22 September 2020
Accepted 24 September 2020

ABSTRACT

Background and purpose The ever-growing availability of imaging led to increasing incidentally discovered unruptured intracranial aneurysms (UIAs). We leveraged machine-learning techniques and advanced statistical methods to provide new insights into rupture intracranial aneurysm (RIA) risks.

Methods We analysed the characteristics of 2505 patients with intracranial aneurysms (IA) discovered between 2016 and 2019. Baseline characteristics, familial history of IA, tobacco and alcohol consumption, pharmacological treatments before the IA diagnosis, cardiovascular risk factors and comorbidities, headaches, allergy and atopy, IA location, absolute IA size and adjusted size ratio (aSR) were analysed with a multivariable logistic regression (MLR) model. A random forest (RF) method globally assessed the risk factors and evaluated the predictive capacity of a multivariate model.

Results Among 994 patients with RIA (39.7%) and 1511 patients with UIA (60.3 %), the MLR showed that IA location appeared to be the most significant factor associated with RIA (OR, 95% CI: internal carotid artery, reference; middle cerebral artery, 2.72, 2.02–3.58; anterior cerebral artery, 4.99, 3.61–6.92; posterior circulation arteries, 6.05, 4.41–8.33). Size and aSR were not significant factors associated with RIA in the MLR model and antiplatelet-treatment intake patients were less likely to have RIA (OR: 0.74; 95% CI: 0.55–0.98). IA location, age, following by aSR were the best predictors of RIA using the RF model.

Conclusions The location of IA is the most consistent parameter associated with RIA. The use of 'artificial intelligence' RF helps to re-evaluate the contribution and selection of each risk factor in the multivariate model.

Check for updates

To cite: Rousseau O, Karakachoff M, Gaignard A, *et al*. *J Neurol Neurosurg Psychiatry* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jnnp-2020-324371

INTRODUCTION

A ruptured intracranial aneurysm (RIA) is a vascular event with a mortality rate as high as 40%.[1] It causes a loss of productive life years similar to that of an ischaemic stroke (the most common type of stroke), and its annual total economic burden in the UK was estimated to be 510 million GBP.[2] In recent years, improved and more widely accessible non-invasive intracranial imaging techniques have led to an increased number of small, incidentally discovered unruptured IAs (UIAs).[3 4] Generally, the overall prevalence of UIAs in the general population is estimated to be 3.2%.[5 6]

Preventive treatment of UIA is a tangible management option given the treatment-related hazards as well as the differential risk of rupture. In the absence of a randomised trial comparing treated patients with conservatively managed patients, treatment of UIAs remains both challenging and controversial, even if better outcomes have been reported for treated patients compared with conservatively managed patients.[7] Consequently, there are neither clear recommendations nor a consensus regarding the optimal management of patients with UIA.[2 8] Past and current studies have suggested that UIAs may be classified as presenting a high or low rupture risk on the basis of their location and size.[10–14] Larger IAs and IAs in the posterior circulation arteries (PCircA) are thought to be related to a higher risk of RIAs.[11 12 13] Furthermore, multiple IAs,[11] female sex,[13] young age,[11 12] history of RIA[14] and cigarette smoking have also been suggested as rupture-predisposing risk factors in various studies. Clinical decisions thus mainly rely on generic risk factors organised by prognostic scores, such as the PHASES (population, hypertension, age, size of the aneurysm, earlier subarachnoid haemorrhage from another aneurysm and site of aneurysm).[15]

The Understanding the Pathophysiology of Intracranial Aneurysm (ICAN) project recruited a collection of patients, including patients with RIAs and those with UIAs, with extensive anatomical and epidemiological characterisation as well as a certified-expert clinical annotation. For instance, the ICAN project prospectively recorded data on tobacco consumption by pack-years, pharmacological treatments before the IA diagnosis and very precise IA location for more than 20 variables in total.[16]

Machine-learning algorithms associated with large-scale computing infrastructure are currently accessible, and they have performed well in classification tasks such as patient stratification.[17] Leveraging both state-of-the-art machine-learning techniques and advanced statistical methods, we provide new insights into IA rupture risks.

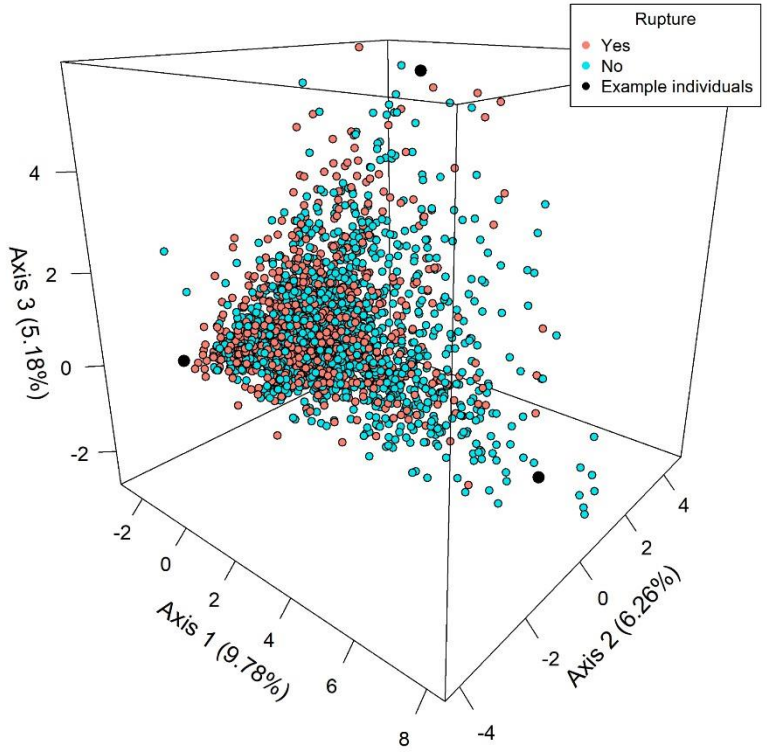
BMJ

Publication sur la méthode Pre-Print : https://www.researchsquare.com/article/rs-1674043/v1

Nantes Université

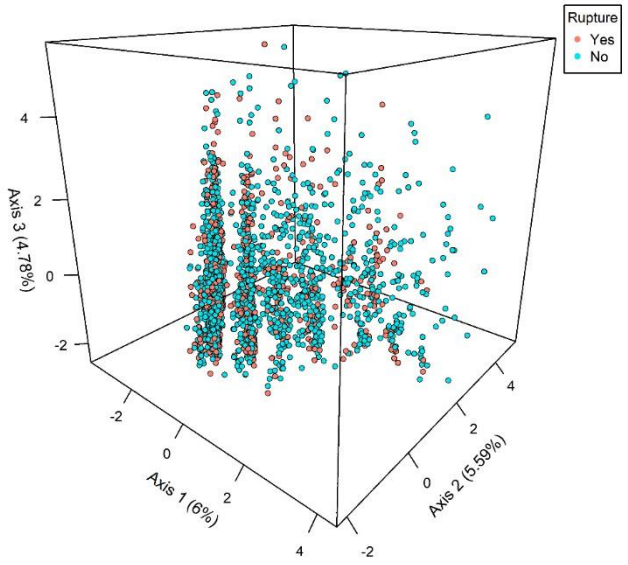
Comparison of FAMD

SENSITIVE DATA (Original Pseudonymous data)



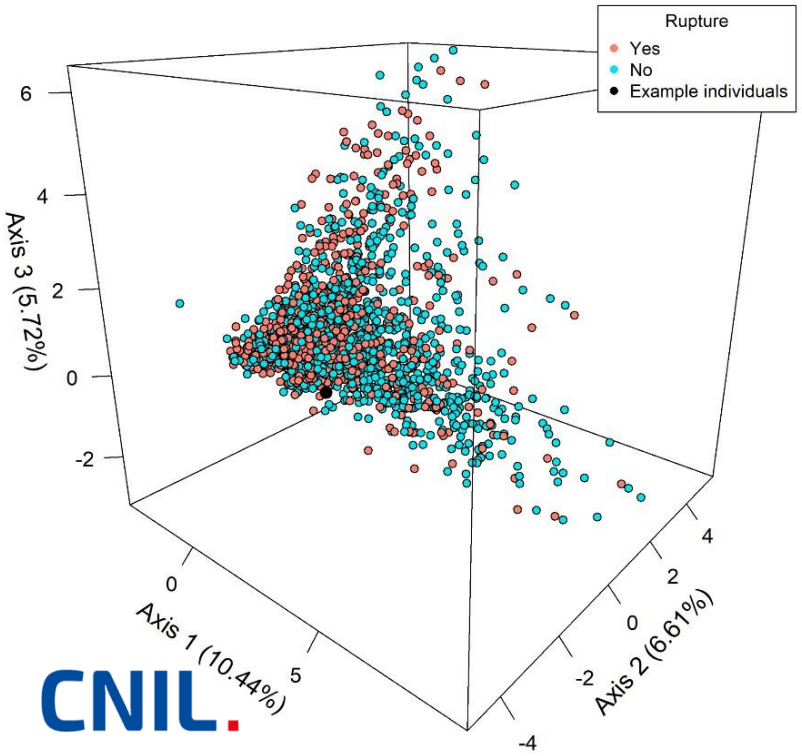
SENSITIVE DATA (Original Pseudonymous data)	
Risk of re-identification	+++
Preparation/curation	+
Informative Value	++++

SIMULATED DATA (mathematically simulated representation of the original dataset)



SIMULATED DATA (mathematically simulated representation of the original dataset)	
Risk of re-identification	0
Preparation	++++
Informative Value	+

AVATARS (synthetic version data the original dataset)



CNIL.

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

Uses Cases : 2 Typical Biomedical (tabular) datasets

• (AIDS) Clinical trial

- The AIDS dataset includes 2139 patients and 26 variables for HIV-infected patients who participated in a clinical trial published in 1996 in the *New England Journal of Medicine*. The clinical trial had four arms and was analyzed by Hammer et al. (1996)[29]. The principal endpoints used were survival and a 50% drop in CD4+ cell counts.

• Wisconsin Breast Cancer Diagnosis (WBCD): prediction issue

- The WBCD dataset comprises 683 observations and 10 variables. It can be downloaded from the University of California Irvine machine-learning repository[30]. The outcome corresponds to the tumor severity: benign (n=444) versus malignant (n=239). The other nine features are built from imaging specific annotations and are graduated from 1–10. Feature selection (F-score computation) and a support vector machine (SVM) were used to predict the severity of a patient's breast cancer diagnosis as per Akay et al. (2009)[31].

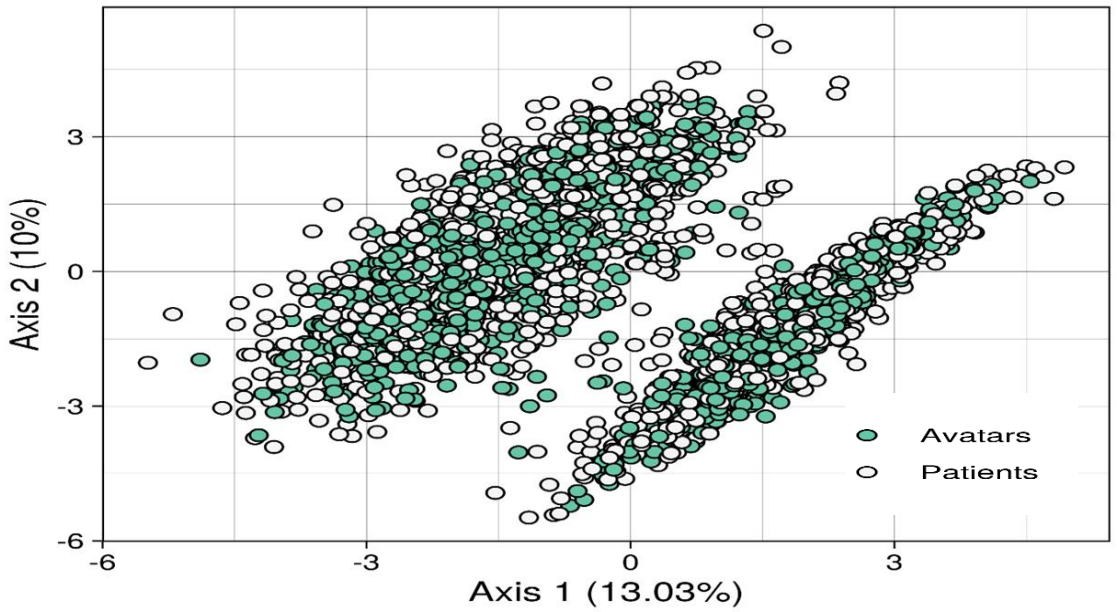
Publication sur la méthode Pre-Print : https://www.researchsquare.com/article/rs-1674043/v1

Results 1: Datasets conservations

- Similar multidimensional representation

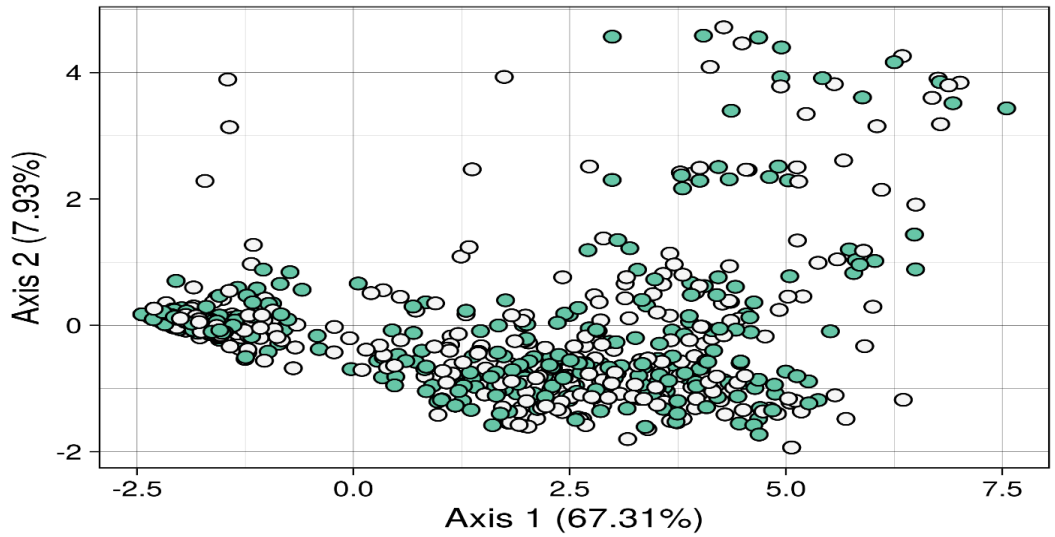
AIDS - Clinical Trial

a



WBCD – cancer prediction

b

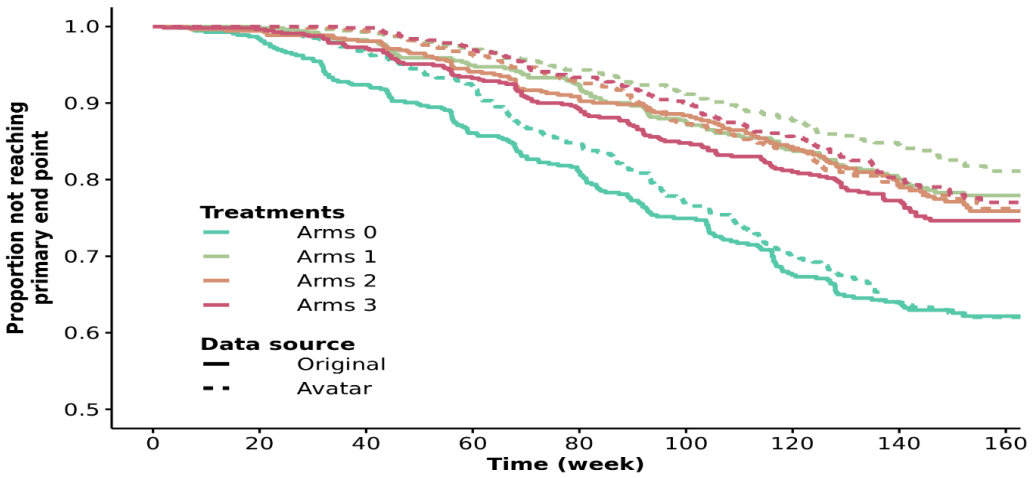


Results 2: Statistics conservations

- Similar results of the main analysis (if exsiting) associated to the data set

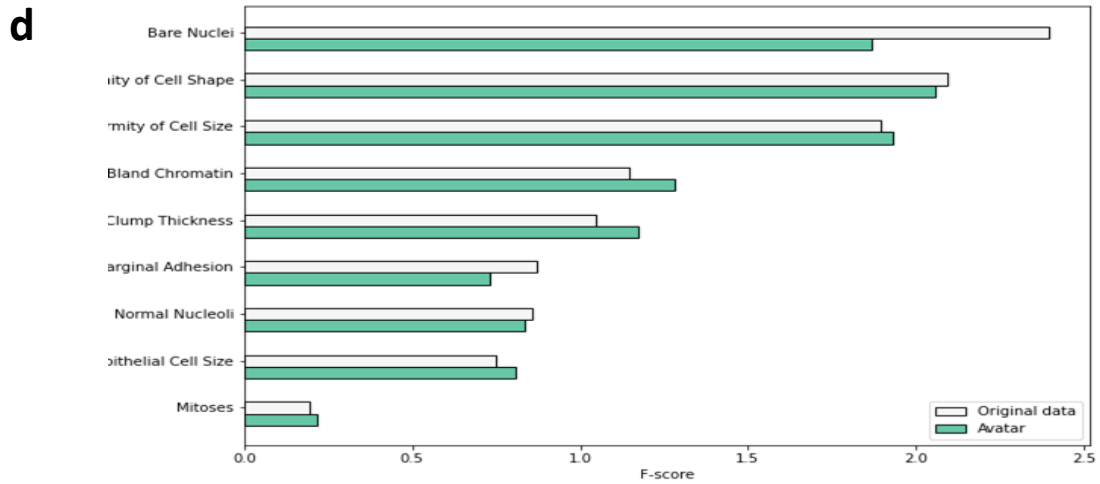
AIDS - Clinical Trial

Avatars p= 1,5 E-9 vs p= 1,2E-8



WBCD – cancer prediction

AUC= 99,84 vs AUC= 99,46



Nantes Université

Results 3: Statistics conservations

- Similar results of the main analysis (if existing) associated to the data set

AIDS - Clinical Trial

Avatars p= 1,5 E-9 vs p= 1,2E-8

	Hazard Ratio	Pr(>|z|)	lower .95	upper .95
Avatar arms1	0.4	<0.001	0.31	0.51
Original arms1	0.49	<0.001	0.39	0.63
Avatar arms2	0.52	<0.001	0.41	0.67
Original arms2	0.52	<0.001	0.41	0.67
Avatar arms3	0.5	<0.001	0.39	0.63
Original arms3	0.59	<0.001	0.47	0.73

WBCD – cancer prediction

AUC= 99,84 vs AUC= 99,46

	avatar	original
acc	99.024390	92.682927
auc	99.186864	99.940312
npv	97.368421	90.140845
ppv	100.000000	94.029851
sens	98.473282	94.736842
spec	100.000000	88.888889

??? Est ce que la valeur statistique des données compte vraiment ?

Peut-être pas en premier…

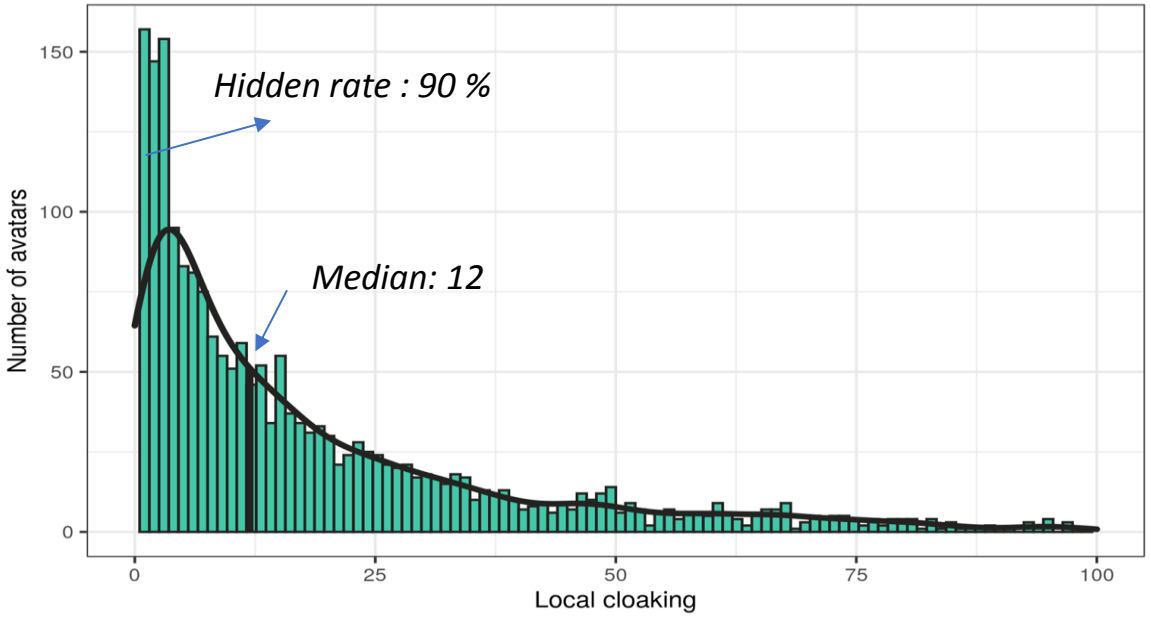
La confiance des patients

«Quand on analyse des données , il n'y a plus de justification à faire courir un risque de ré-identification aux patients. »

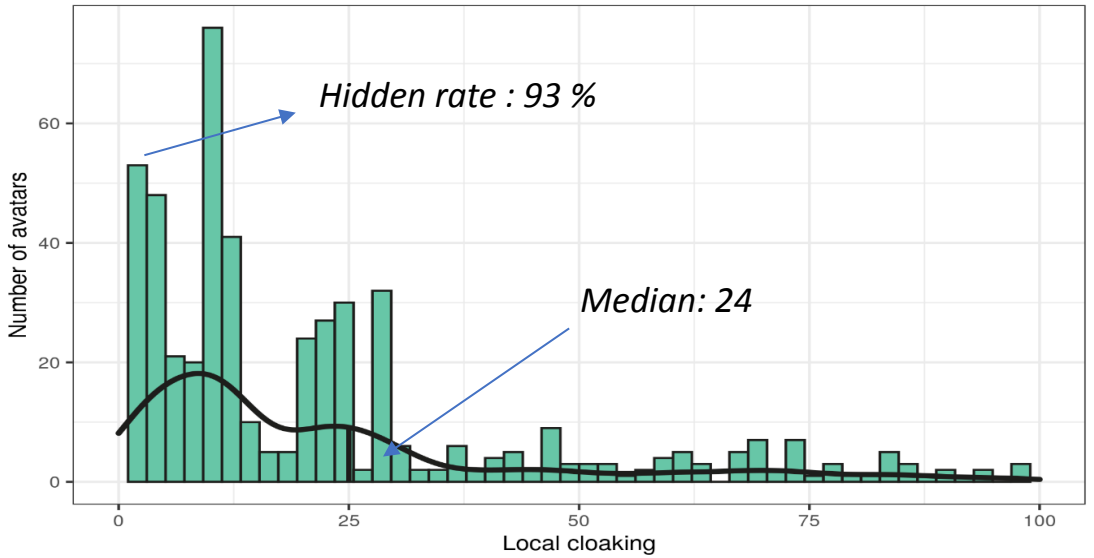
Results I : Privacy matters most

- How well protected sensitive observations are ?

12 avatars in average
– 10 % with their avatars at their side



25 avatars in average –
5,7% with their avatars as the closest



Nantes Université

Les CHU un “carrefour des données” de Santé ?

Articulation entre modes d’accès nationaux – accompagnement local

- Circuit d’information individuelle systématisée
- Proximité de la production des données : « En santé, un data scientist est d’abord un expert du contexte dans lequel naît la donnée avant d’être un expert des méthodes de traitement de ces données”

Médiation par un tiers-expert

Structure locale labellisée

- « les données parleraient d’elles-mêmes »
- “Jamais seul face aux données”
- HUB local - Clinique des données – Centres de Données Cliniques

Un réservoir d’innovation…

- TAL, données synthétiques, apprentissage fédéré, chiffrement homomorphique, Pilotage par la multidata

Enjeu de transformation – nouvelle épidémiologie de données.

On parle très vite en millions…

Données en « vie réelle » - qui requièrent plus de méthodes

Nantes Université

L'innovation autour de la donnée (massive)

(1) La Clinique des donnée et le réseau de CDC; (2) Objet connecté; (3) Données synthétiques

Prof. Pierre-Antoine Gourraud, Nantes Université & CHU

15 Juin 2022, Faculté de Pharmacie

HUGO
Hopitaux Universitaires Grand Ouest

- **COI :**

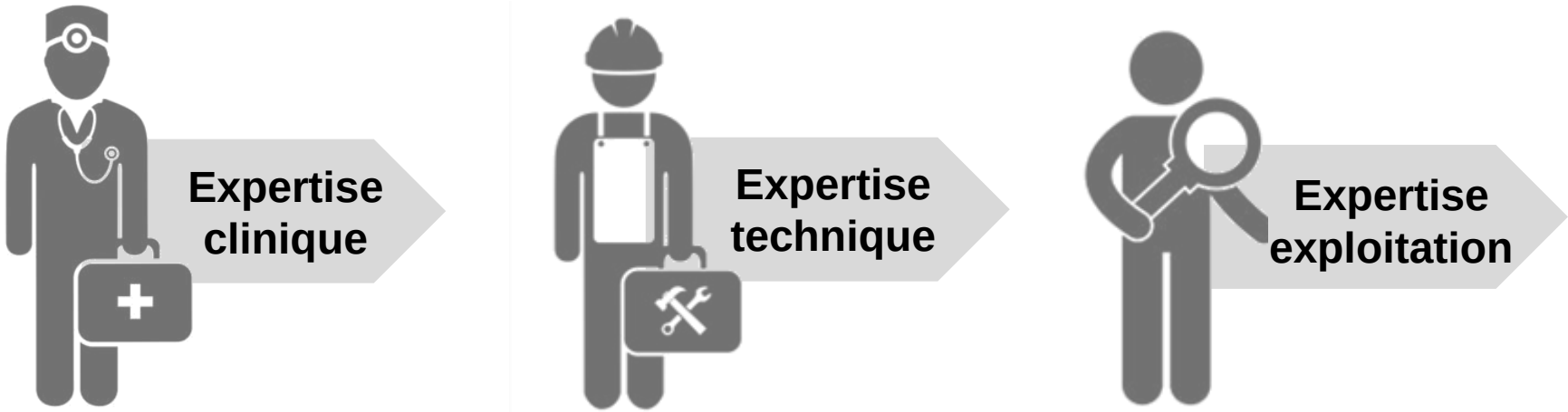
PA Gourraud is the founder of Methodomics (2008) and the co-founder of Big data Santé (2018). He consults for major pharmaceutical companies, all of which are handled through academic pipelines (AstraZeneca, Biogen, Boston Scientific, Cook, Docaposte, Edimark, Ellipses, Elsevier, Methodomics, Merck, Mérieux, Sanofi-Genzyme, Octopize). PA Gourraud is volunteer board member at AXA mutual insurance company (2021). He has no prescription activity with either drugs or devices.

« Verrou » des données

→ « **Bonnes** » **données** : un préalable à de bonnes analyses

→ Les données, c'est l'**expertise** du domaine

Compréhension et qualité des données sont un enjeu croissant à mesure que la disponibilité technique est résolue par les outils numériques



Les données méritent une « clinique » au CHU de Nantes

Nantes Université

Circuit de sollicitation

1. **Définir la question scientifique : investigateur**

2. **Formuler/Officialiser le besoin d'accompagnement sur le Portail recherche** (http://ehopappprd:8084/portail-recherche) : investigateur
   -> orientation de la demande
3. **Déterminer le niveau d'accompagnement en fonction du besoin** : investigateur et CdD

Comme un service Clinique avec des patiients
Accompagner, former : Open Space
- Jamais seul face aux données -

Montage

- Appui méthodologique
- Réponse AO/AAP (dont dénombrement eHOP)

Réalisation

- Screening / datamart
- Constitution d'une cohorte
- Appariement multi-bases
- Analyse des données

Valorisation

- Communication lors de congrès
- Publication

Synthetic Data will get popular

(1) Structural similarity

- (i.e., the same granularity) : same statistical unit
- *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.

(3) Subjective assessment:

- Neither experts, nor trained algorithms can distinguish synthetic data from original data.