8e ÉDITION

JOURNÉES DU **GFCO 2022**

Biomarqueurs et analyses moléculaires en oncologie

Avec la participation scientifique du GFCO

PLÉNIÈRE - REAL-WORLD DATA
Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins

Pr. Pierre-Antoine Gourraud, Nantes

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

Nantes Université, INSERM, CRT2I , Nantes, France

pierre-antoine.gourraud@univ-nantes.fr

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

Inserm

Institut national de la santé et de la recherche médicale

LIENS D'INTÉRÊT

- **COI :**
- ***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 the founder of Methodomics (2008) and the co-founder of Big data Santé (2018).***
- **PA Gourraud is volunteer elected board member at AXA mutual insurance company (2021).**
- **He has no prescription activity with either drugs or devices.**

Introduction et Plan

- « Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins »
- **Intro et Détour par la génétique (épidémiologique)**
  - 3 concepts « approches agnostique », secondaires et intentionnalité
  - Une nouvelle activité hospitalière « la clinique des données »
- **« Parcours de soins - RWE » Quelles Données ? SNDS et EDS**
- **« Prendre soin des données » : Les entrepôts de de données de santé, les clinique des données dans les établissements**
  - Gouvernance légalité technique
  - EXEMPLES
- **Conclusion : Carrefour de données de Santé**

“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.

“Re-penser les données de Santé”

LE FIGARO lundi 12 septembre 2022

SANTÉ 19

Repenser la production des données de santé

PROFESSEUR PIERRE-ANTOINE GOURRAUD
• Professeur de la faculté de médecine de Nantes Université, praticien hospitalier

DATA Les données de santé sont parmi les données personnelles les plus sensibles. Le déploiement de la plateforme française des données de santé, le Health Data Hub, pose donc des questions inédites de souveraineté technologique, et la possible création d'un espace européen des données de santé en posera d'autres. Pour y répondre, nous devons repenser ce qu'est un « producteur de données de santé », et reconnaître que ces données sont « coproduites » par soignants et patients.

Dans le numérique, les données sont souvent réduites à un problème informatique dont la solution est technique. Mais les patients seraient-ils de simples « producteurs » de données de santé, comme un arbre porterait du « fructus » ? Ces données seraient-elles assimilables à des biens matériels, et réduites à leur valeur de transaction commerciale ? Cette vision des données, souvent prise pour du pragmatisme qui subordonne le politique à l'économique, induit en erreur tant les décideurs politiques que le grand public. La nature profondes données de santé est plus complexe que cela.

D'abord, d'un point de vue légal, les données de santé ne sont pas des biens patrimoniaux. Elles restent intimement liées à une personne et chaque patient a le droit de s'opposer à ce que les données qui le concernent soient utilisées dans certains buts à l'exclusion d'autres, par exemple pour la recherche. Mais elles ne leur appartiennent pas au même titre qu'un objet qu'ils posséderaient et dont ils pourraient disposer librement, il est donc plus juste de parler de données « qui concernent les patients » plutôt que de « leurs » données. Dans l'usage courant, l'adjectif possessif (« nos » données) tend à faire des données un objet marchand, sous l'influence du droit anglo-saxon de la data. Mais pour des données personnelles, et encore plus si elles touchent à notre santé, c'est plutôt le champ du droit à l'image qui est pertinent. Un patient exerce un contrôle sur les utilisations des données émanant de lui, comme il le ferait pour une photographie, sans avoir à justifier l'exercice de son droit d'opposition.

Parler de « producteur de données de santé » est de surcroît injuste pour la communauté des soignants. Tous, à leur échelle, des médecins aux ingénieurs biomédicaux en passant par les aides-soignants, contribuent à « coproduire » les données de santé des patients qui résultent de leurs observations cliniques, de leurs analyses en imagerie ou en biologie médicale, et de leurs jugements. Il faut donc bien parler a minima de « coproduction » de ces données entre les patients et les soignants. Si les patients sont concernés, les soignants le sont aussi car ces données touchent à l'intimité de l'exercice de l'art de la médecine.

Le déploiement de la plateforme française des données de santé, le Health Data Hub, pose des questions de souveraineté technologique. ALAIN JOCARD/AFP

Responsabilité régalienne

Patients et soignants sont donc coconcernés, cocontributeurs, coacteurs et coauteurs de la santé. Tout autant objets que sujets des données de santé, leurs rôles ne sont jamais vraiment passifs. Ainsi si les « producteurs » de données de santé produisent, c'est plus au sens musical qu'agricole ! Cette production musicale, il nous la faut « danser » entre mouvements et contributions multiples, comme l'évoquait Nietzsche à propos des vérités dont les données sont la figure de proue du XXIe siècle.

Réduire la création de valeur à partir des données de santé à un enjeu de technologie informatique est donc une double erreur. Si l'enjeu technique est surestimé, il prend le pas sur les enjeux de gouvernance des données, chez – et entre – les acteurs de la santé, à l'hôpital comme en libéral. La technique demeure un moyen au service d'une fin, sans chercher une revanche sur une supposée toute-puissance médicale.

À l'inverse, si l'enjeu technique est sous-estimé on risque de traiter ces données de santé avec légèreté, et d'oublier le vieux serment d'Hippocrate qui habite l'inconscient de chaque soignant, qui, « admis dans l'intimité des foyers taira les secrets qui lui seront confiés ». À contrario, le citoyen conscient des usages secondaires possibles avec des données de santé le concernant, leur donne une seconde vie et exprime sa solidarité avec tous les malades.

La crise du Covid nous a rappelé que la santé est une responsabilité régalienne, incluant les données de santé. Alors cessons d'être naïfs et donnons à la France un « réseau d'éditeurs de données de santé ». Du Health Data Hub à chaque acteur de santé (au plus proche du soin et de l'émergence de données), seule une organisation décentralisée permettra d'accroître la confiance dans les usages des données, et de conjuguer les intelligences dans leur exploitation pour mieux évaluer et faire évoluer les services de santé au service de la société. ■

Le Pr Pierre-Antonie Gourraud déclare ne pas avoir de lien d'intérêt en rapport avec le sujet traité. Il est le fondateur en 2008 de Methodomics (www.methodomics.com) et le cofondateur de Wedata en 2018 (www.wedata.science). 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, Edimark, Ellipses, Haevier, Methodomics, Merck, Mérieux, Sanofi-Genzyme, WeData). Il est administrateur bénévole des mutuelles d'assurances Avsa (2021). Il n'a aucune activité de prescription de médicaments ou de dispositifs médicaux.

LE PLAISIR DES LIVRES

- ***« Patients et soignants sont donc co-concernés, co-contributeurs, co-acteurs et co-auteurs de la santé. »***
- ***« Ainsi si les « producteurs » de données de santé produisent, c’est plus au sens musical qu’agricole ! »***

PA Gourraud

Le Figaro 12 Septembre 2022

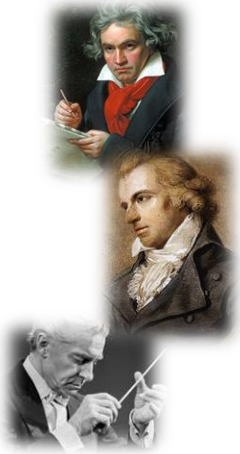
Intégration **des données** pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins

Les données massives en santé

Avec la participation scientifique du GFCO

Introduction to the true nature of (medical) Data

- **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 …**



The "Ode to Joy"

- *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…

Nantes Université

Définitions : Données Massives en Santé

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 :*

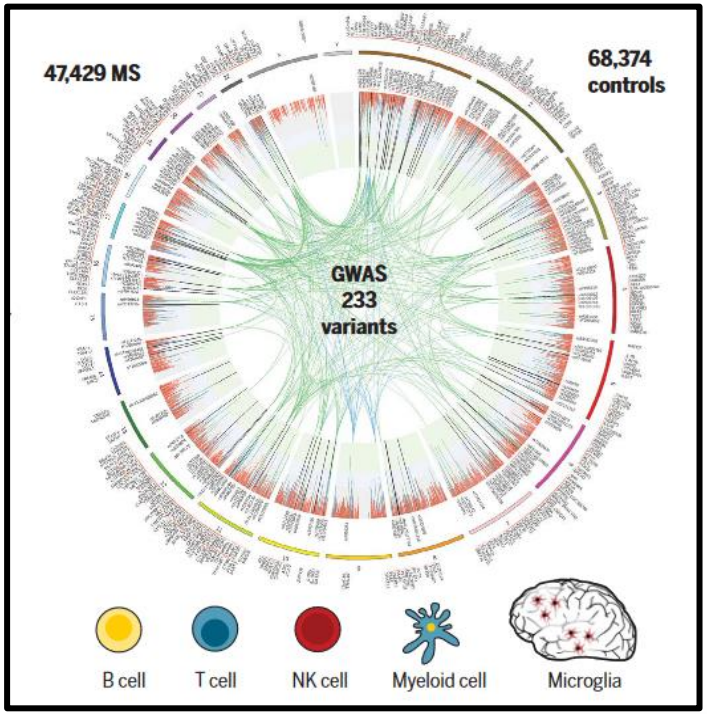
Données structurées	Données non structurées
• Valeurs biologiques (hémoglobine, glycémie…) • Données génétiques • PMSI • …	• Compte-rendu : Texte « brut » -> TALN • Données de signal (ECG) • Images (scanner…) • …

Intégration des données pour l'émergence de **nouveaux concepts** et l'amélioration du parcours de soins

Détour par le Génétique

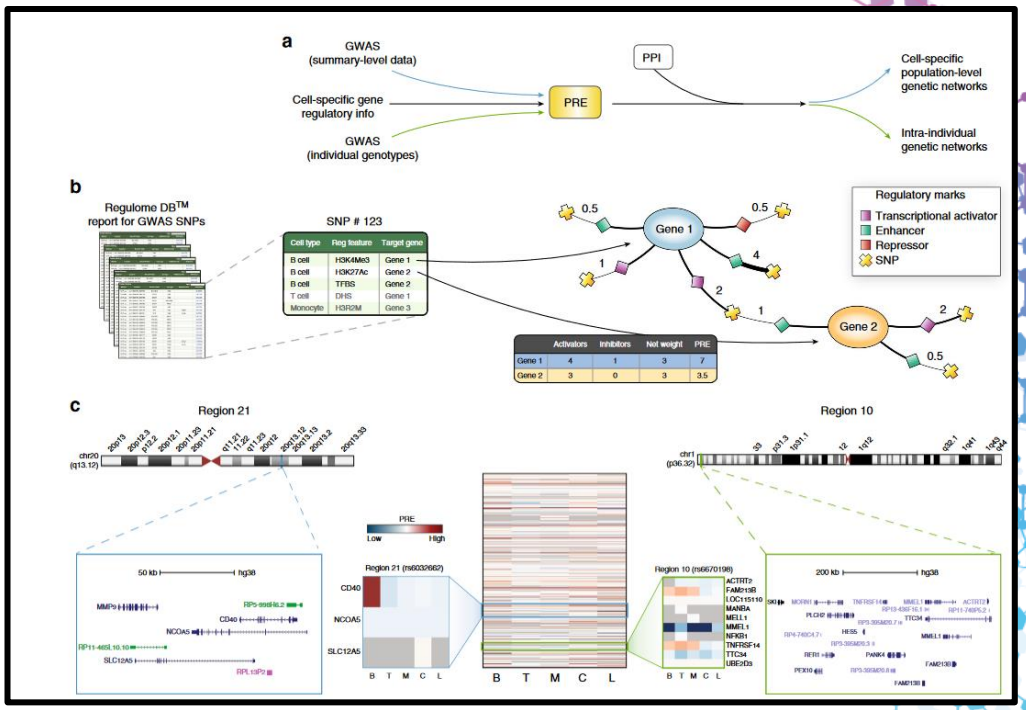
Avec la participation scientifique du GFCO

The International Multiple Sclerosis Genetics Consortium (IMSGC)



Science 2019: Genomic map

We analyzed genetic data of 47,429 MS and 68,374 control subjects and established a reference map of the genetic architecture of MS. This map includes 32 variants within the extended Major Histocompatibility Complex (MHC), 200 autosomal susceptibility variants outside the MHC, and one chromosome X variant.

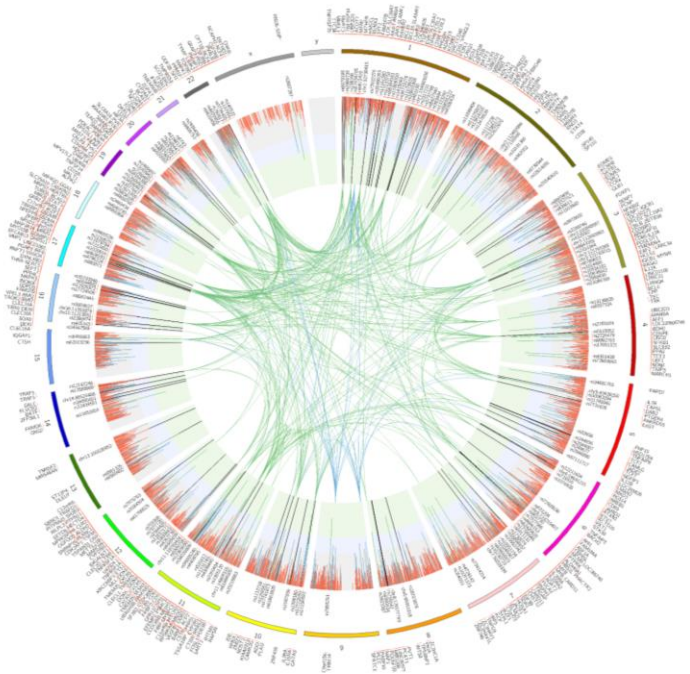


Nature Communications 2019: Pathways

We conducted a cell-specific pathway analysis of the latest GWAS in MS. Our analysis identifies pan immune cell as well as cell-specific susceptibility genes in T cells, B cells and monocytes. Additionally, genotype-level data from 2,370 patients and 412 controls were used to compute intra-individual and cell-specific susceptibility pathways that offer a biological interpretation of the individual genetic risk to MS.

Update : 200+ Common Variants

- ***"The Multiple Sclerosis Genomic Map: Role of peripheral immune cells and resident microglia in susceptibility"***
  - **IMSGC et al. 2019 Science**
- **Large-scale comparison**
  - 47,351 multiple sclerosis (MS) subjects
  - 68,284 control subjects
- **> 200 MS associated genomic regions**
  - 200 autosomal susceptibility variants
  - 1 chromosome X variant,
  - 32 independent associations in the extended MHC
  - Prioritize up 551 potentially associated MS genes



Establish a reference map of the 200+ genetic architecture of MS

Imsgc et al. Science 2019

3 “nouveaux “ Concepts vus par le détour de la génétique

1. *Rupture de la massification*
2. *« Approches agnostique » dans les données*
3. *Développements d’usage secondaires*

Intentionnalité des données...

…

Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du **parcours de soins**

Accès aux données de Parcours de Soins…

Avec la participation scientifique du GFCO

Les données de parcours de Soin

- **Soins individuels**
  - Le soin prodigué au patient est la raison d'être du partage de données entre soignants
  - Portabilité (interopérabilité)

- **L'analyse du parcours de soins**
  - Inter-établissements
    - Interopérabilité (comparabilité)
  - Via prisme des données médico-administratives
    - Centralisées et centralisables
  - Entrepôt de données de GHT, Coordination entre EDS de type Ouest Data Hub

Nantes Université in West of France

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

HUGO Hopitaux Universitaires Grand Ouest

2016 : From Local Biomedical Data Warehouses

- **Development et optimization**:

  1- Data Governance

  2- Legal framework

  3- Technical Issue.

UNIVERSITÉ DE RENNES 1

Pr Marc CUGGIA

Laboratoire Traitement du Signal et de l'Image LTSI

enovacom filiale d'Orange Business Services

ehop

Technical Deployment - public- private partnership

enovacom filiale d'Orange Business Services

RiCDC

Nantes Université

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 ehop	SNDS Système national des données de santé **+ 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** biobank uk CONSTANCES 1000 Genomes Project HGDP Human Genome Diversity Project

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

Inserm
Institut national de la santé et de la recherche médicale

Nantes Université

CENTRE HOSPITALIER
UNIVERSITAIRE DE NANTES

Les données :
c'est le vecteur de transformation du XXIème 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

Inserm
Institut national
de la santé et de la recherche médicale

Une gouvernance Data qui doit être un catalyseur

Clinique des Données
(MOA)

Traitements des données

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

DRI - Resp. Juridique
(Légal et financier)

Règlementaire Recherche, partenariats, juridique

Lien DPO

DSN – Resp . Technique
(MOE)

Intégration, stockage, sécurisation

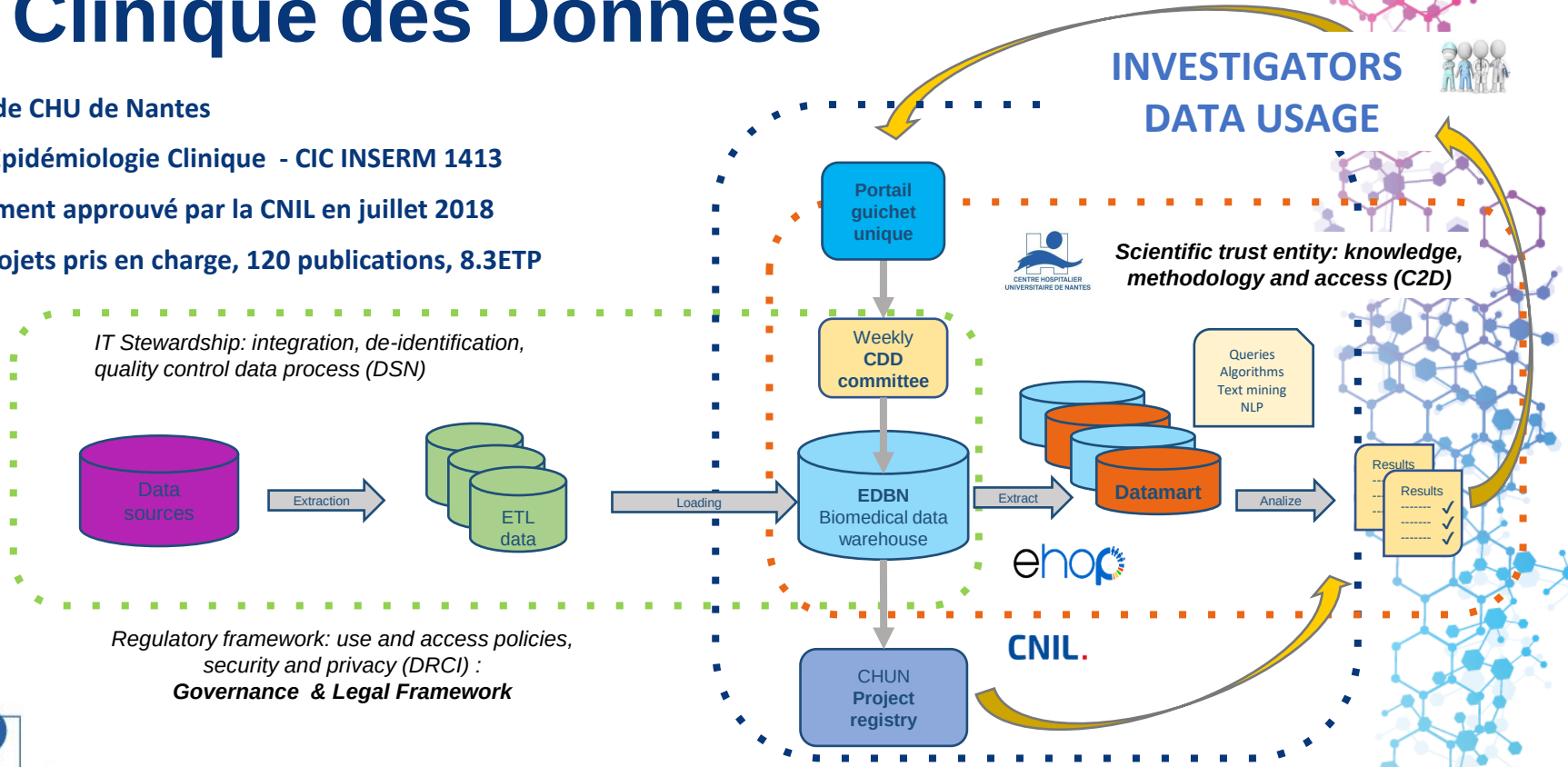
Cellule épidémiologie clinique du CIC 1413

Service de Santé Publique – PHU11

Expertise = assistance à …
MOE , MOA, Légal et financier
Voir à « l'urbanisation »

La Clinique des Données

- Service de CHU de Nantes
- Cellule Epidémiologie Clinique - CIC INSERM 1413
- Déploiement approuvé par la CNIL en juillet 2018
- > 400 projets pris en charge, 120 publications, 8.3ETP



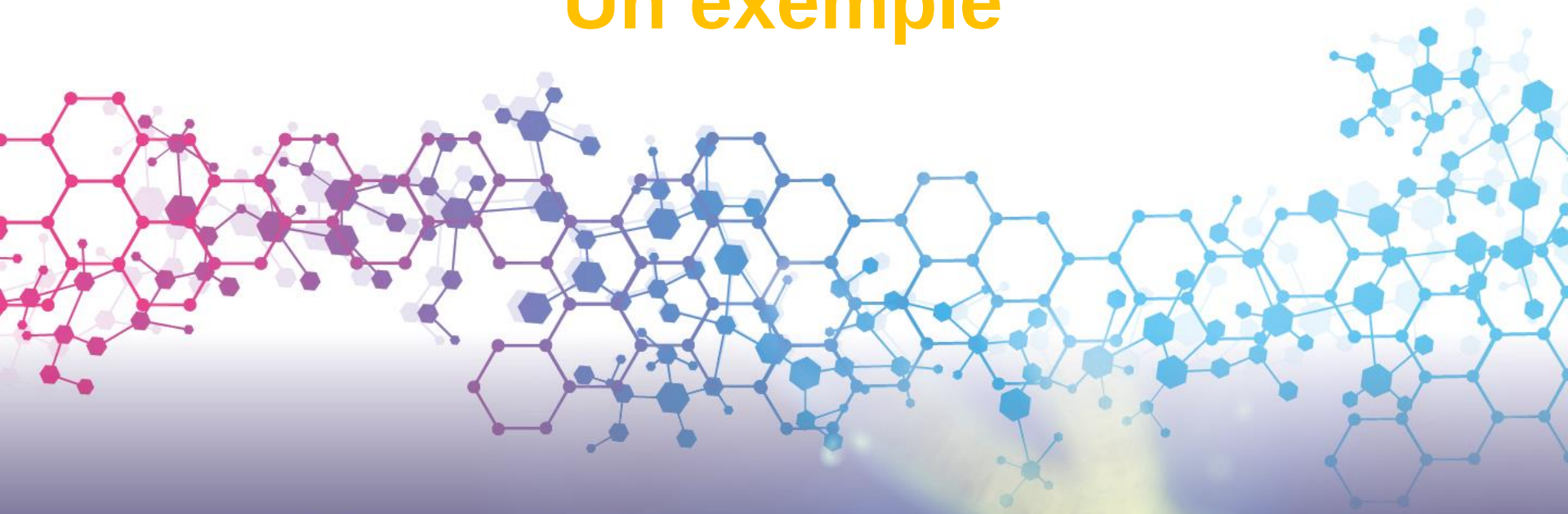
CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

Inserm
Institut national de la santé et de la recherche médicale

Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins

Un exemple



Avec la participation scientifique du GFCO

Illustrations

	RESPONSABILITÉ	
	OUI	NON
INTENTIONNALITÉ OUI		
INTENTIONNALITÉ NON	X	

Projet SEVASAR

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

Illustrations

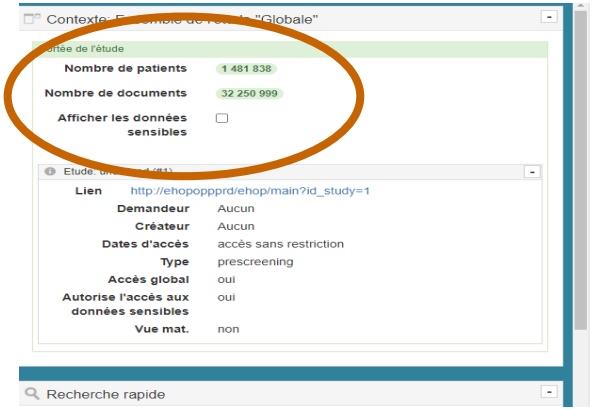
Projet SEVASAR

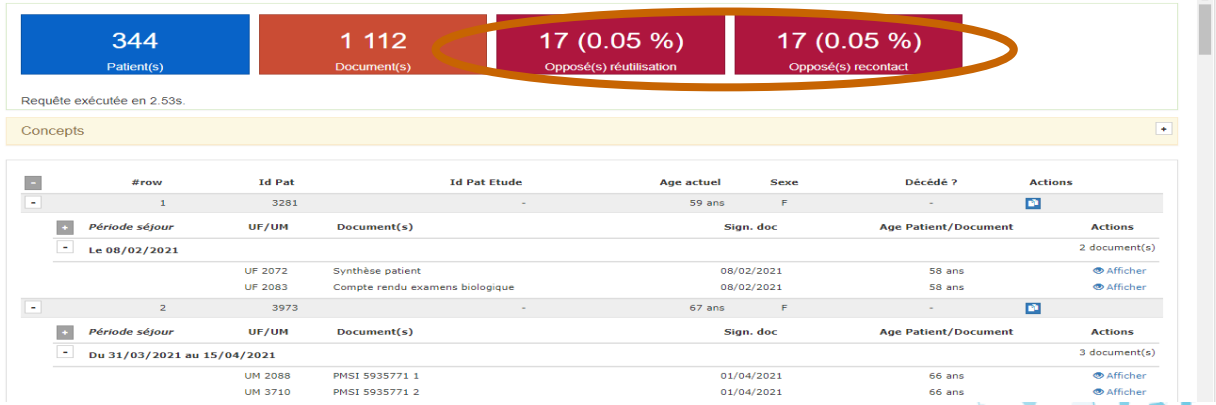
Identification des patients atteints du variant anglais du COVID-19

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

	RESPONSABILITÉ	
	OUI	NON
INTENTIONNALITÉ OUI		
INTENTIONNALITÉ NON	X	

3. Recherche eHOP par interface graphique : résultats





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)

Illustrations

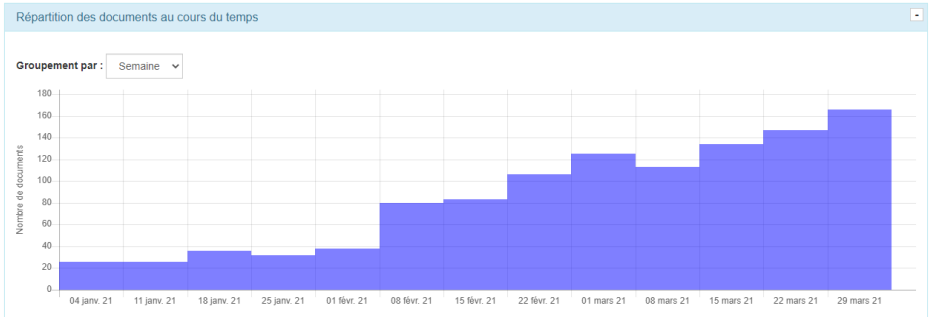
Projet SEVASAR

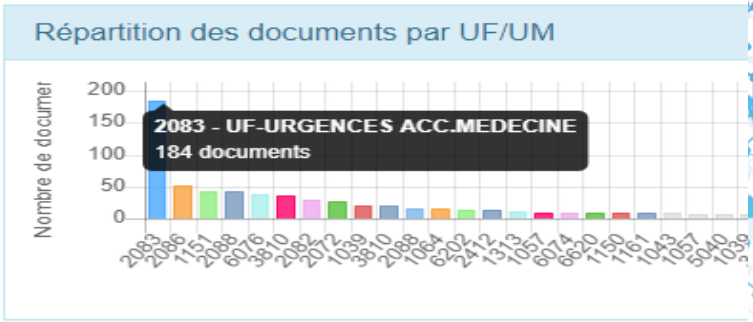
Identification des patients atteints du variant anglais du COVID-19

Dr P. Le Turnier
Service Maladies infectieuses et tropicales

INTENTIONNALITÉ \ RESPONSABILITÉ	OUI	NON
OUI		
NON	X	

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





Les “carrefours de 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

PERSPECTIVE

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

PLÉNIÈRE - REAL-WORLD DATA
Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins

Pr. Pierre-Antoine Gourraud, Nantes

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

Nantes Université, INSERM, CRT2I , Nantes, France

pierre-antoine.gourraud@univ-nantes.fr

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

Inserm

Institut national de la santé et de la recherche médicale

Intégration des données pour l'émergence de nouveaux concepts et l'amélioration du parcours de soins

Un bonus

Avec la participation scientifique du GFCO

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.

« L'analyse de données personnelles dans un but d'analyse non personnelle devient un enjeu éthique »

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**.

Using synthetic Data in biomedical data warehouse

- **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

OCTOPIZE
MIMETHIK 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 Michelozzi ,[9] Gaultier Marnat,[10] Anne-Clémence Vion,[2] Gervaise Loirand,[4] Hubert Desal,[11] Richard Redon,[2] Pierre-Antoine Gourraud,[1] Romain Bourcier ,[11] The ICAN Investigators

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

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

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.[8 9] 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 15] Furthermore, multiple IAs,[11] female sex,[12] young age,[12 13] 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).[8]

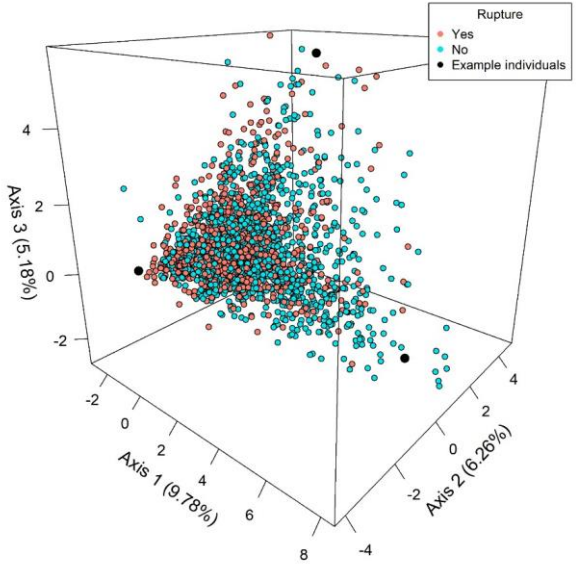
The Understanding the Pathophysiology of Intracranial Aneurysms (ICAN) project encouraged a collection of patients, including patients with RIAs and those with UIAs, with extensive anatomical and epidemiological characterisation as well as a certified-expert clinical annotation. For instance, the ICAN project prospectively recorded data on tobacco consumption by pack-years, pharmacological treatments before the IA diagnosis and very precise IA location for more than 29 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.

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

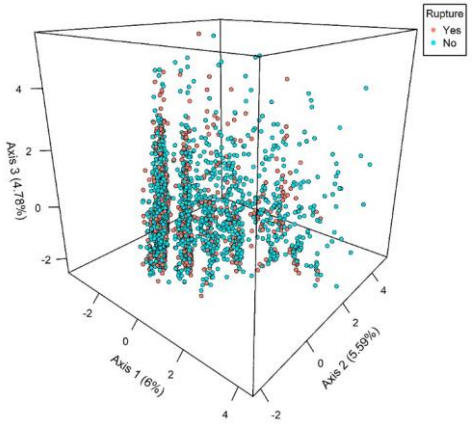
Comparison of FAMD

SENSITIVE DATA
(Original Pseudonymous data)



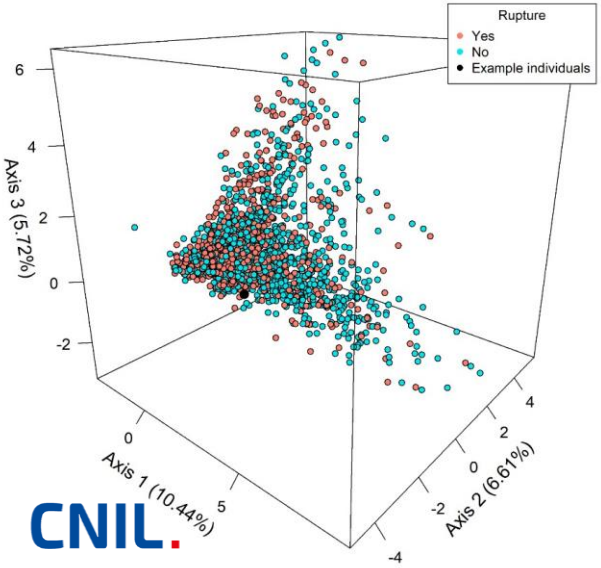
SENSITIVE DATA (Original Pseudonymous data)	
Risk of re-identification	+++
Preparation/curation	+
Informative Value	++++

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].

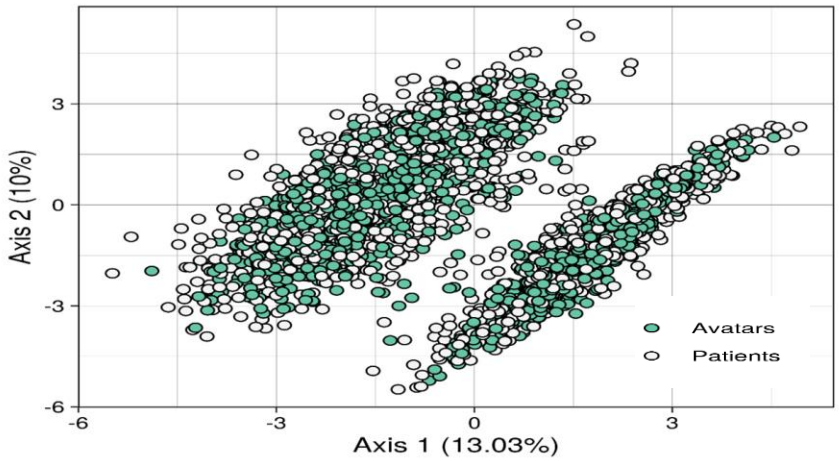
Publication sur la méthode Pre-Print : https://www.researchsquare.com/article/rs-1674043/v1

Results 1: Datasets conservations

- Similar multidimensional representation

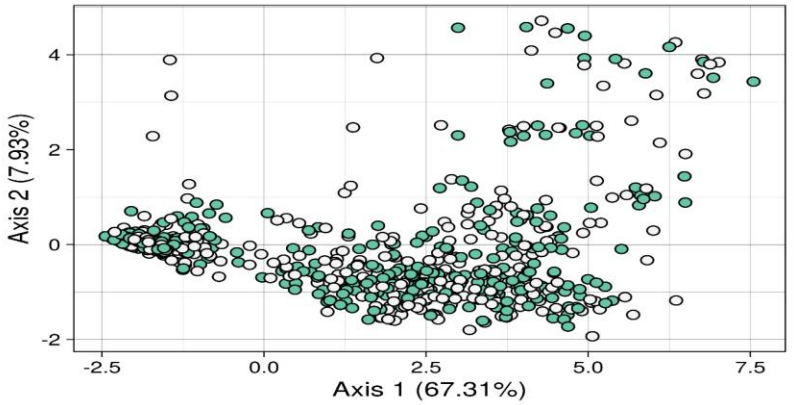
AIDS - Clinical Trial

a



WBCD – cancer prediction

b



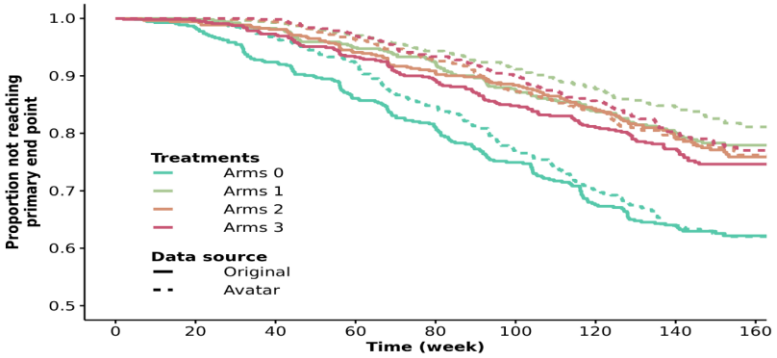
Results 2: Statistics conservations

- 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

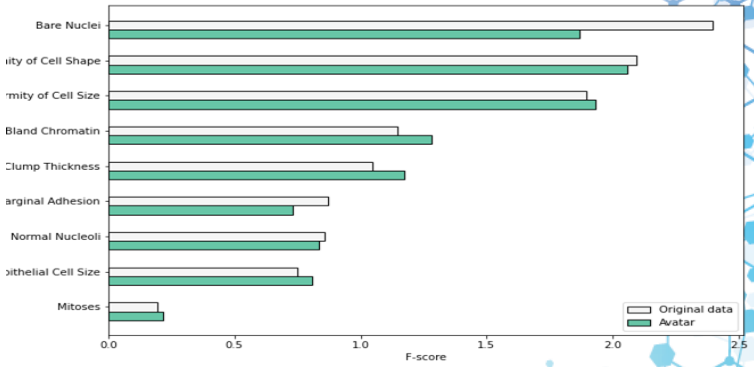
c



WBCD – cancer prediction

AUC= 99,84 vs AUC= 99,46

d



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

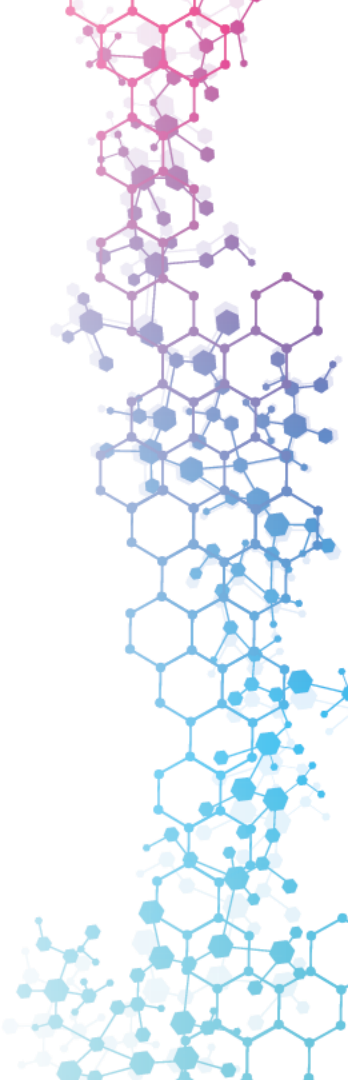
WBCD – cancer prediction

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…



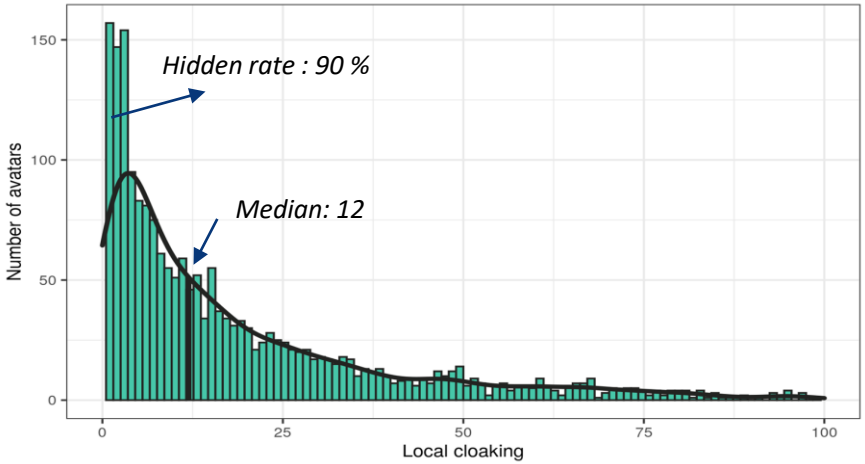
La confiance des patients

«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

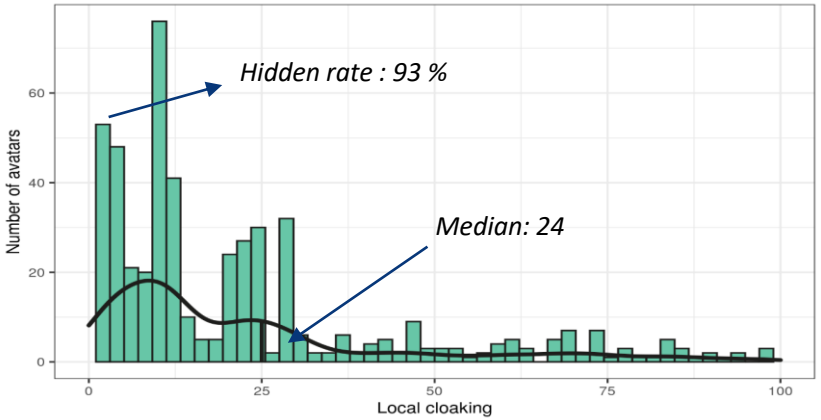
- How well protected sensitive observations are ?

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



AIDS - Clinical Trial

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



WBCD – cancer prediction

Merci de votre attention