DU 14 AU 16 MARS
ESPACE MAYENNE
LAVAL
2023

WEST DATA FESTIVAL

DE L'INTELLIGENCE POUR VOS DONNÉES

14 mars 10h30 – 11h

Introduction et état des lieux :
Simplifier la prise en charge d'un patient sénior grâce à l'utilisation des données

Quelles sont les ressources et les besoins en matière de recueil de données dans le parcours de soins d'un patient ? Quels sont les outils existant aujourd'hui et quels sont les enjeux ? Mieux comprendre la gestion des données dans le secteur des soins de santé.

WWW.WESTDATAFESTIVAL.FR

Par Pierre-Antoine Gourraud, PU-PH de la faculté de médecine de Nantes Université

Organisé par :

Co-organisé avec :

Les partenaires du West Data Festival 2023 :

LMT est soutenue par :

Enjeux de données de santé

- ***“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

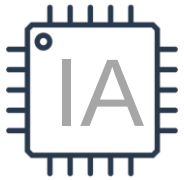
Les moyens existent de libérer l’exploitation des données tout en protégeant la confidentialité, notamment pour le croisement et le partage des fichiers de vaccinations et de tests, expliquent, dans une tribune au « Monde », Yann Coatanlem, président du club de réflexion Praxis et Pierre-Antoine Gourraud,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

Enjeux de données de santé : Définitions : Données Massives (1/2)

= « **Big data** » ⇨ 

1. Rupture dans la **Volumétrie**

2. **Non intentionnalité** des données *= collecte dans le cadre de la pratique courante ou usage secondaire*

3. **Variété** dans la typologie des données
   Définition par les sources et par les structures

Enjeux de données de santé : Définitions : Données Massives (2/2)

> *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…) • …

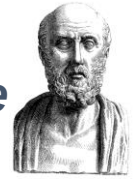
Enjeu dans le rapport au patient : la confidentialité

- **Les données identifiantes sont celles qui vont permettre la ré-identification d'un individu.**
  - *Directement* identifiantes, car pouvant être uniques à un individu : prénom, nom, numéro de sécurité sociale (aussi appelé NIR)
    - = Dé-identifier les données : (attention à ne pas dire anonymiser)
  - *Indirectement identifiantes* parce que permettant, par recoupement avec d'autres informations, d'identifier un individu de façon unique.
    - = Données pseudonymisées – on peut recréer un lien (un « pseudonyme »)
- **Anonymiser**
  - Démontrer l'impossibilité de revenir à l'individu à l'origine des données
- **Confidentialité des patients et … des soignants.**
  - Principe de parcimonie.

Sécurité des données de santé, Un enjeu des SI de Santé; D.I.C.T

“Primum Non Nocere

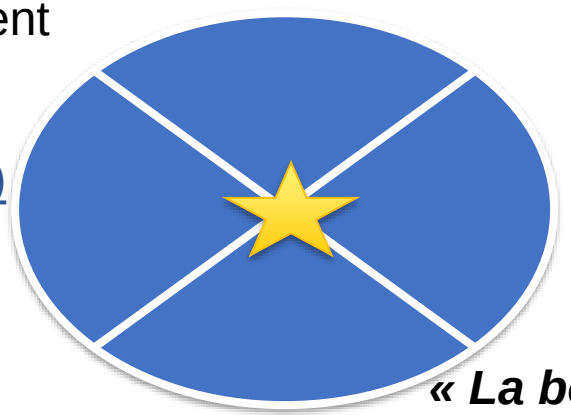
Hippocrates c. 460 – c. 370 BC



Disponibilité

« A la demande »

- Données ou services : accessibles rapidement et régulièrement



Confidentialité

« Pas open »

- Utilisateur : autorisé à avoir accès aux informations (notions de droits ou permissions).
- Patient : Respect de la vie privée : Données Personnelles & Données de Santé

Traçabilité (Preuve / Authenticité)

« Rendre des comptes »

- Identito-vigilance = I*dentification (Qui) et authentification (mot de passe/secret)*.
- Confiance dans les relations d'échange (Soigné soignants établissements) .

Intégrité

« La bonne donnée, le bon service pour un usage »

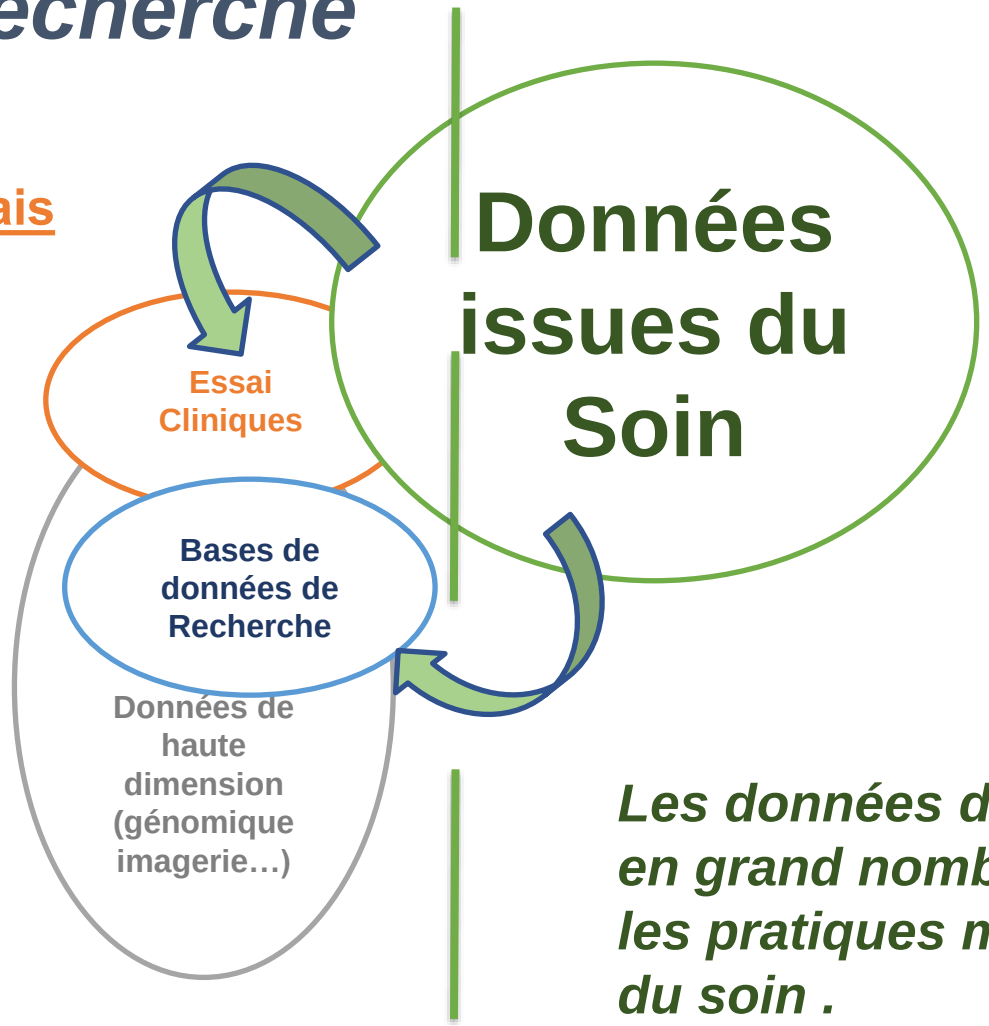
- Altération des données et erreur des requêtes
- Archivage // mémoire des actes et des enregistrements

Un glissement dans les données disponibles à des fins de recherche

Données d'essais cliniques

Données cohorte/registre

Données de recherche biomédicale



Données DMP/DPI/EMR/EHR

- *Usage non-intentionnel à des fins de recherche*
- « Population réelle » « dans mon CHU »
- Lacunaires
- Réponse à un ensemble de questions
- Taille ouverte mais limites pratique
- Coûts faibles – Réflexion a posteriori

Les données de soins deviennent réutilisables en grand nombre pour évaluer et faire évoluer les pratiques mais leur intentionnalité reste celle du soin .

« Big Data » = Rupture quantitative /Massification dans les données disponibles

Pôle Santé UFR Médecine

DEFLP-HI DEep Learning Proposal for Healthcare & Innovation NeXT AI & Health data

Inserm Institut national de la santé et de la recherche médicale

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

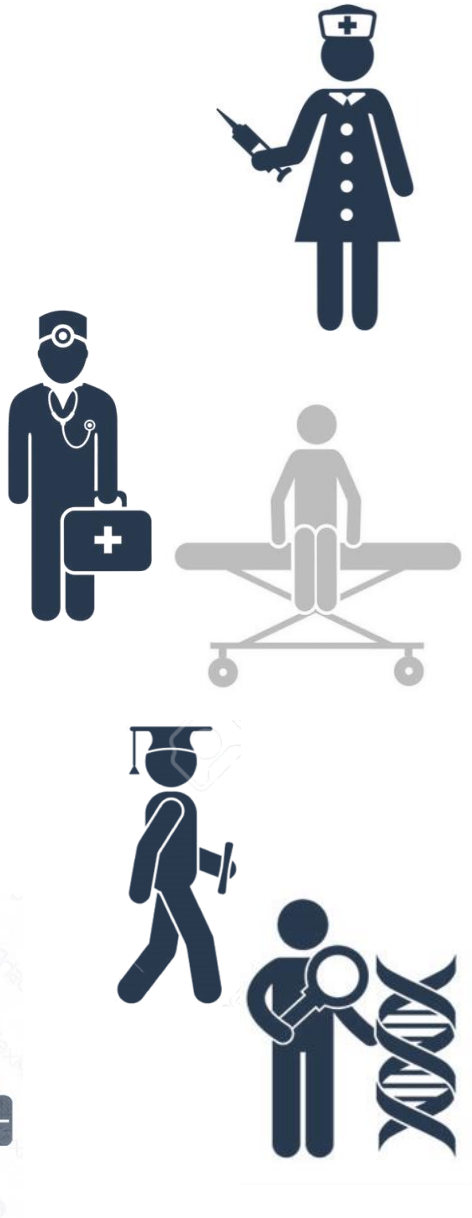
Nantes Université

Nantes Université

La valeur des données de santé résulte d'une chaine de contributions multiples qui en accroissent la valeur d'usage.

Les 7 usages des données massives – « Big Data »

1. **Usages multi-niveaux dans le pilotage médico-administratif du système de soins**
2. **Usages de ciblage sous-groupe (public et privé)**
3. **Irruptions du Big Data dans la relation soigné-soignant (Pôle quantitatif objectif)**
   - Prescrire (drugs and data)
   - Comparer (Soli-data-rity)
   - Anticiper (Prédire et mesure l'incertitude des scenarii thérapeutiques)
4. **Le malade seul face aux données (des données trop accessibles et non évaluées ?)**
   - Continuité santé bien-être.
   - Irruptions du Big Data dans l'émergence d'un bio-pouvoir santéiste (hors cadre prof santé)
5. **Big data dans la détection d'événements anormaux**
   - Phase IV, pharmacovigilance, BMR, stérilisation, infectio-vigilance
6. **Usage en recherche : "Épidémiologie du XXIème siècle"**
   - Données de recherché enrichies pas des données de soins (Biais)
7. **Enjeux pour la formation des professionnels de Santé**
   - Usage et interaction conditionnées par ces usages des données



Plateformes des données de santé

- Accès facilité aux données de santé de sources multiples (hôpitaux, ville, pharmacie, biologie…)
- Environnements sécurisés
- Valorisation du « patrimoine » de données de santé
- Hub national et hubs locaux

Entrepôt(s) de données biomédical
Etablissement région National
HUGO

OUEST DATA HUB

HEALTH DATA HUB

4,6 millions patients

112 millions clinical documents

ehop

A single IT for 8 institutions at the crossroad of healthcare system for 11 millions people

1.3 billions structured data

5+ millions Visits

Plateformes des données de santé

SNDS
Système national des données de santé

SNIRAM
- 1,2 Milliard feuilles soins/an
- Remboursement sécu (médicaments, examens para-cliniques, AT…)

PMSI
- Données hospitalisation (11 millions de séjours /an)
- Motifs d'admission
- Diagnostics (CIM-10)
- Actes CCAM

CépiDc
- Décès (600 000 /an)
- Causes de décès

ALD

Age, sexe, commune résidence, cmu

Pseudo-nymisées
Chaînées par le NIR

Données synthétiques

- Pas de donnée clinique
- Pas de résultat para clinique
- Pas de donnée sur les médicaments non remboursés
- Pas de données sur les revenus

Le digital s'inscrit dans un élargissement de la relation soigné-soignant .

- **En fait, ce n'est pas nouveau ….avec le digital**

Depuis le vieux stéthoscope jusqu'au jeune appareil à résonance magnétique nucléaire, en passant par la radiographie, la scannographie, l'échographie, la scientificité de l'acte médical éclate dans la substitution symbolique du laboratoire d'examens au cabinet de consultation. Parallèlement l'échelle du plan de représentation des phénomènes pathologiques se transforme, de l'organe à la cellule, de la cellule à la molécule.

Le statut épistémologique de la médecine par Georges Canguilhem Source: *History and Philosophy of the Life Sciences* Vol. 10, Supplement: Medicine and Epistemology. Health, Disease and Transformation of Knowledge Perugia, Italy. 17-20 April, 1985 (1988), pp. 15-29

- **Relation Soigné – Soignant + un tiers …**
  - Renforcement d'un troisième pole de plus en plus digital
    - Tout en étant technique biologique complexe

→ Rôle de médiateur renforcé du soignant

  - Dans le rapport du soigné au pathologique triplement incarné par : "sa maladie", "les connaissances biotechniques médicales", "le soignant"
  - Pas si paradoxal que ça.


Vous ne guérirez peut-être pas, mais vous survivrez à travers vos données !
Tesson

Valoriser les données en mettant à jour le statut vital

Base de données des personnes décédées de l'INSEE est nominative et open-data…

Applications:

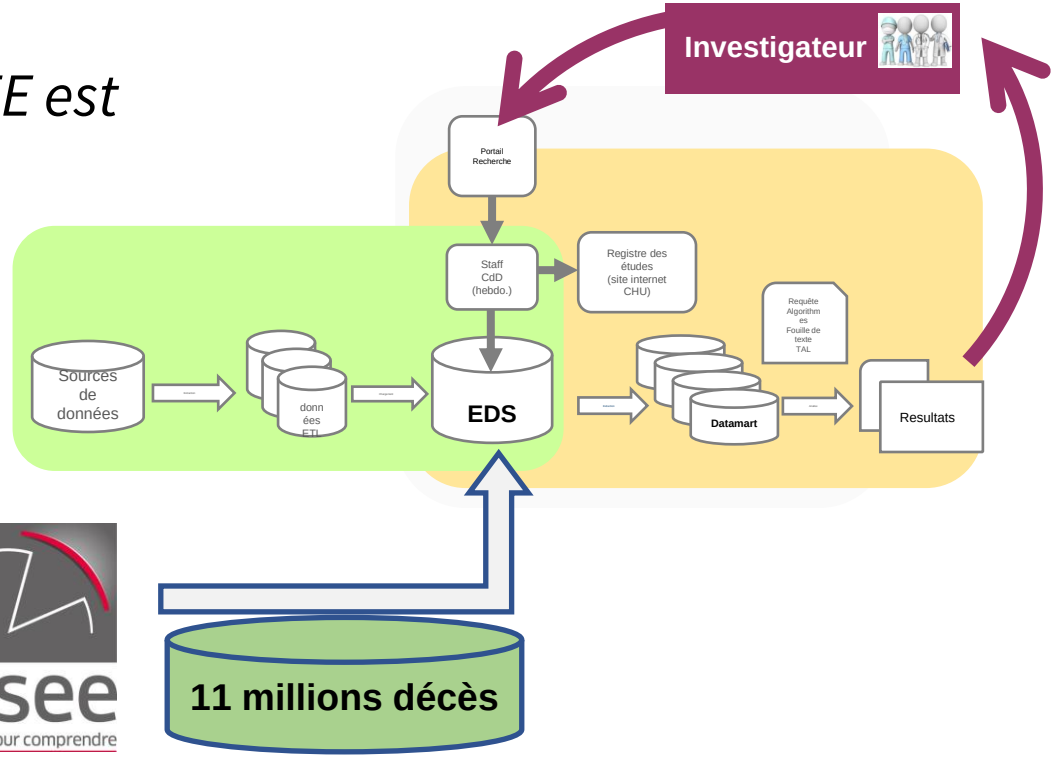
- Décès: critère de jugement important dans les études cliniques
- Information administrative importante

Challenges:

- Appariement optimal

Results:

1. 205,698 (10.4%) morts identifiés
2. Sensibilité : 93.3% [92.8–93.9]
3. ~10% d'amélioration avec un algorithme "intelligent"



Insee
Mesurer pour comprendre

JMIR Publications
Advancing Digital Health & Open Science

Guardiolle et al. 2022
Paper : https://medinform.jmir.org/2022/11/e36711
Code source : https://gitlab.com/ricdc/insee-deces

Au-delà de la publication…

- **Publication :** https://medinform.jmir.org/2022/11/e36711
- **Large-scale matching algorithm for linking biomedical dat warehouse records with the national mortality database i France**
  - Collaboration NU CHU - LS2N Informatique
  - Collaboration CHU Nantes Rennes Lille
- **La place centrale de la validation des innovations en santé**
- **Faire le choix de l'ouverture**
  - Github: https://gitlab.com/ricdc/insee-deces
  - Permettre que d'autres s'en saisissent (HSJSA , ICO)

JMIR MEDICAL INFORMATICS Guardiolle et al

Original Paper

Linking Biomedical Data Warehouse Records With the National Mortality Database in France: Large-scale Matching Algorithm

Vianney Guardiolle[1], MSc, MD; Adrien Bazoge[1,2], MSc; Emmanuel Morin[2], PhD; Béatrice Daille[2], PhD; Delphine Toublant[1], MS; Guillaume Bouzillé[3], MD, PhD; Youenn Merel[3], MSc; Morgane Pierre-Jean[3], PhD; Alexandre Filiot[4], MSc; Marc Cuggia[3], MD, PhD; Matthieu Wargny[1], MD, MSc; Antoine Lamer[5], PhD; Pierre-Antoine Gourraud[1,6], PhD

[1]CHU de Nantes, INSERM CIC 1413, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, Nantes, France
[2]Le Laboratoire des Sciences du Numérique de Nantes, Université de Nantes, Nantes, France
[3]Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, Rennes, France
[4]CHU Lille, Integration Center of the Lille University Hospital for Data Exploration (INCLUDE), Lille, France
[5]Univ. Lille, CHU Lille, ULR 2694, METRICS: Évaluation des Technologies de santé et des Pratiques médicales, Lille, France
[6]Université de Nantes, CHU de Nantes, INSERM, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ATIP-Avenir, Nantes, France

Corresponding Author:
Antoine Lamer, PhD
Univ. Lille, CHU Lille, ULR 2694
METRICS: Évaluation des Technologies de santé et des Pratiques médicales
42 Rue Paul Duez
Lille, 59000
France
Phone: 33 665760587
Email: antoine.lamer@univ-lille.fr

Abstract

Background: Often missing from or uncertain in a biomedical data warehouse (BDW), vital status after discharge is central to the value of a BDW in medical research. The French National Mortality Database (FNMD) offers open-source nominative records of every death. Matching large-scale BDWs records with the FNMD combines multiple challenges: absence of unique common identifiers between the 2 databases, names changing over life, clerical errors, and the exponential growth of the number of comparisons to compute.

Objective: We aimed to develop a new algorithm for matching BDW records to the FNMD and evaluated its performance.

Methods: We developed a deterministic algorithm based on advanced data cleaning and knowledge of the naming system and the Damerau-Levenshtein distance (DLD). The algorithm's performance was independently assessed using BDW data of 3 university hospitals: Lille, Nantes, and Rennes. Specificity was evaluated with living patients on January 1, 2016 (ie, patients with at least 1 hospital encounter before and after this date). Sensitivity was evaluated with patients recorded as deceased between January 1, 2001, and December 31, 2020. The DLD-based algorithm was compared to a direct matching algorithm with minimal data cleaning as a reference.

Results: All centers combined, sensitivity was 11% higher for the DLD-based algorithm (93.3%, 95% CI 92.8-93.9) than for the direct algorithm (82.7%, 95% CI 81.8-83.6; *P*<.001). Sensitivity was superior for men at 2 centers (Nantes: 87%, 95% CI 85.1-89 vs 83.6%, 95% CI 81.4-85.8; *P*=.006; Rennes: 98.6%, 95% CI 98.1-99.2 vs 96%, 95% CI 94.9-97.1; *P*<.001) and for patients born in France at all centers (Nantes: 85.8%, 95% CI 84.3-87.3 vs 74.6%, 95% CI 72.8-76.4; *P*<.001). The DLD-based algorithm revealed significant differences in sensitivity among centers (Nantes, 85.3% vs Lille and Rennes, 97.3%, *P*<.001). Specificity was >98% in all subgroups. Our algorithm matched tens of millions of death records from BDWs, with parallel computing capabilities and low RAM requirements. We used the Inseehop open-source R script for this measurement.

Conclusions: Overall, sensitivity/recall was 11% higher using the DLD-based algorithm than that using the direct algorithm. This shows the importance of advanced data cleaning and knowledge of a naming system through DLD use. Statistically significant differences in sensitivity between groups could be found and must be considered when performing an analysis to avoid differential biases. Our algorithm, originally conceived for linking a BDW with the FNMD, can be used to match any large-scale databases. While matching operations using names are considered sensitive computational operations, the Inseehop package released here

https://medinform.jmir.org/2022/10/e36711/ JMIR Med Inform 2022 | vol. 10 | iss. 10 | e36711 | p. 1
(page number not for citation purposes)

SL•FO

Guardiolle et al. 2022 JMIR

Le plus intéressant n’est jamais la technique…

La vision fixiste et patrimoniale de la donnée est une impasse

Hiérarchie et intrication des normes :

- **Gouvernance :**

  DICT : Quand enrichir la donnée ?

- **Légal :**

  Une personne réputée vivante est un sujet de droit, un personne réputée décédée beaucoup moins …

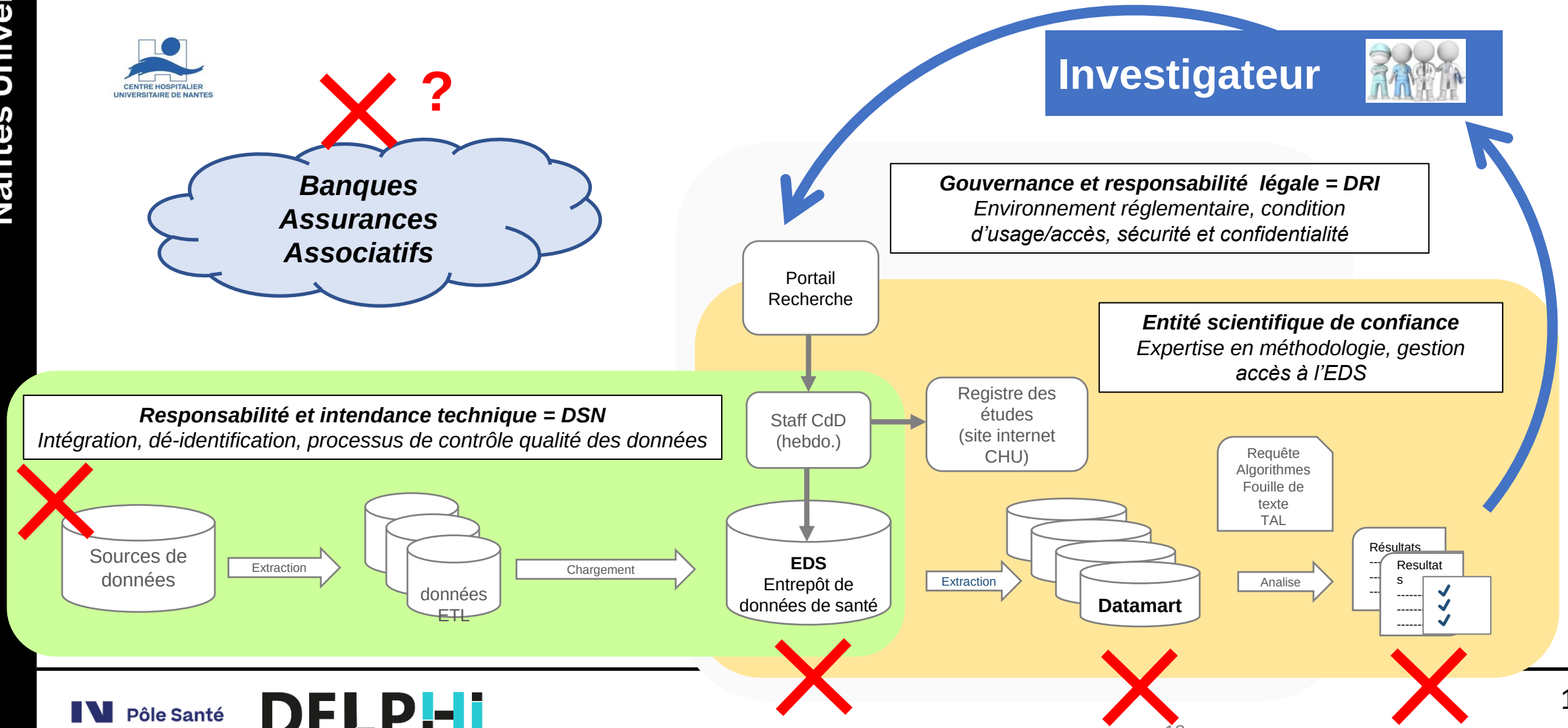
- **Technique :**

  Mise jour, algorithme apprenant, appariement nominatif, traitement massif, algo open source etc etc

- **Citoyen:**

  La massification, du tableau en liège de la mairie de mon village au monde devenu village sur internet …

Le plus intéressant n'est jamais la technique…



"Re penser les données de Santé"

« 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

Inserm
Institut national de la santé et de la recherche médicale

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

Nantes Université

SANTÉ

Repenser la production des données de santé


SERVICE DE PRESSE

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 profonde des 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.


ALAMY VIA REUTERS

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

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 ». A 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-Antoine 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, Elsevier, Methodomics, Merck, Mérieux, Sanofi-Genzyme, WeData). Il est administrateur bénévole des mutuelles d'assurances Axa (2021). Il n'a aucune activité de prescription de médicaments ou de dispositifs médicaux.

LE PLAISIR DES LIVRES

Pôle Santé
UFR Médecine

DELPHI
DEep Learning Proposal for Healthcare & Innovation
NeXT Nantes Université AI & Health data

DU 14 AU 16 MARS
ESPACE MAYENNE
LAVAL
2023

WEST DATA FESTIVAL

DE L'INTELLIGENCE POUR VOS DONNÉES

14 mars 10h30 – 11h

Introduction et état des lieux :
Simplifier la prise en charge d'un patient sénior grâce à l'utilisation des données

Quelles sont les ressources et les besoins en matière de recueil de données dans le parcours de soins d'un patient ? Quels sont les outils existant aujourd'hui et quels sont les enjeux ? Mieux comprendre la gestion des données dans le secteur des soins de santé.

WWW.WESTDATAFESTIVAL.FR

Par Pierre-Antoine Gourraud, PU-PH de la faculté de médecine de Nantes Université

Organisé par :

Co-organisé avec :

Les partenaires du West Data Festival 2023 :

LMT est soutenue par :