







## DONNEES DE SANTE QUALITE DES DONNEES



P. STACCINI (UCA, Nice), A. BENIS (Holon Institute of Technology, Israël), J. GROSJEAN & SJ. DARMONI (URN, Rouen)

### **PLAN**

- Position du sujet
- Données de santé
- Qualité des données
- Données ouvertes



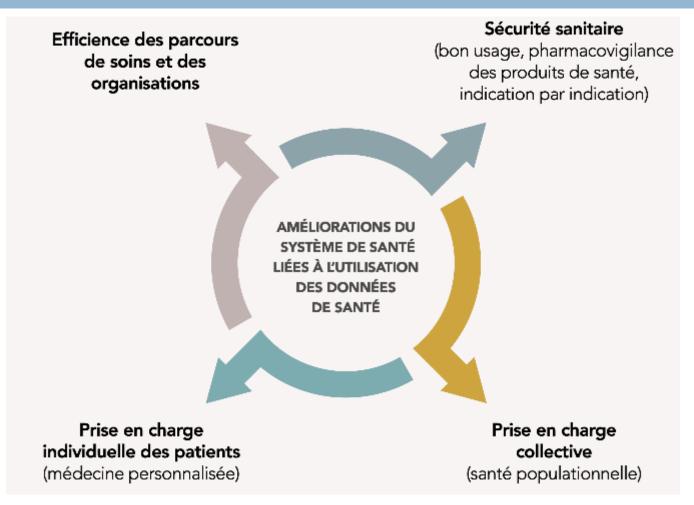
# Position du problème

#### Constats

- Aucun système de santé ne peut fonctionner sans informations de qualité.
- Aujourd'hui, beaucoup de pays ne recensent ni les naissances ni les décès et n'enregistrent pas non plus d'autres informations importantes sur la santé de la population.
- Les données sanitaires sont souvent fragmentaires.
- Plus des deux tiers de la population mondiale vit dans des pays qui n'établissent pas de statistiques fiables sur la mortalité par âge, par sexe et par cause de décès – l'un des indicateurs sanitaires les plus importants pour comprendre quelles sont les priorités d'un pays en termes de santé.
- La moitié seulement des pays ont rapporté à l'OMS des données sur les causes de décès en 2014, et plus de 100 pays ne disposent pas de systèmes fiables pour enregistrer les naissances et les décès.
- De nombreux pays n'ont pas de données de qualité sur leurs personnels de santé ou pour leur système de financement de la santé.
- En raison du manque de données, il est plus difficile de prendre de bonnes décisions sur l'allocation des ressources pour améliorer la santé et aider les gens à vivre plus longtemps et en meilleure santé, et à être plus productifs.



### Enjeux des usages de la donnée

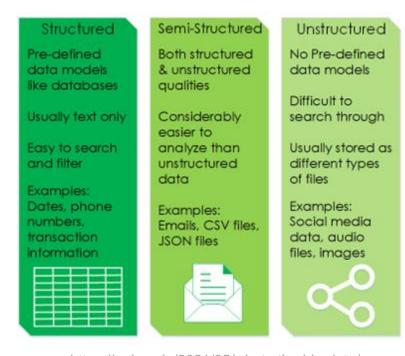




# Données de santé

#### Données massives (Big data), définitions

### Types de données



https://mdaca.io/2021/05/whats-the-big-data/

Structurées :
Biologie (SIL),
Données PMSI (diagnostics, actes)
Prescription, délivrance, administration des medicaments
Dispositifs Médicaux (prescription)
Non structurées : Les textes
Vidéos
Sons,
Semi-Structurées :
Ensemble de données structurées et non

structurées



## Attributs et objets Qu'est-ce qu'une donnée?

- Collection de d'objets de données et leurs attributs
- Un attribut est une propriété ou une caractéristique d'un objet
  - Exemples : couleur des yeux d'une personne, temperature de l'air, etc.
  - Attribut = variable, champ, caractéristique, dimension, ou métadonnée
- Une collection d'attributs décrivent un objet
  - Objet = enregistrement, point, cas, échantillon, entité, ou instance (patient...)



#### **Attributs** Refund Marital **Taxable** Cheat Status Income 125K No Yes Single 2 No Married 100K No 3 No Single 70K No 4 Yes Married 120K No 5 Divorced Yes No 95K 6 No Married 60K No Yes 220K Divorced No 8 No Single 85K Yes 9 No Married 75K No 10 No Single 90K Yes



Credits: Introduction to Data Mining by Tan, Steinbach, Kumar (2004)

### Données, information, connaissance

- Une affectation d'un attribut pour une donnée spécifique génère une information
  - E.g. Température = 38°... Celsius, pas F (US), pas K (International)
- Un règle conditionnelle et formellement décrite d'une information constitue la connaissance
  - E.g. si Température > 38°C alors fièvre
- Des triplets sémantiques représentent aussi la connaissance
  - Deux concepts liés par une relation spécifique
  - Un chat IS-A (est-un) mammifère (voir cours terminologies)



# From Data to Wisdom... or the Big Data Holy Grail The DIKW model



#### Changing the rules

- New policies, Policies updates
- New missions, Missions updates
- New Objectives, Objectives updates
- etc...

#### Knowledge

#### **Human interpretation**

- Insights
- Conclusions
- Opinions
- etc...

#### Information

#### Models reflecting real-world behavior

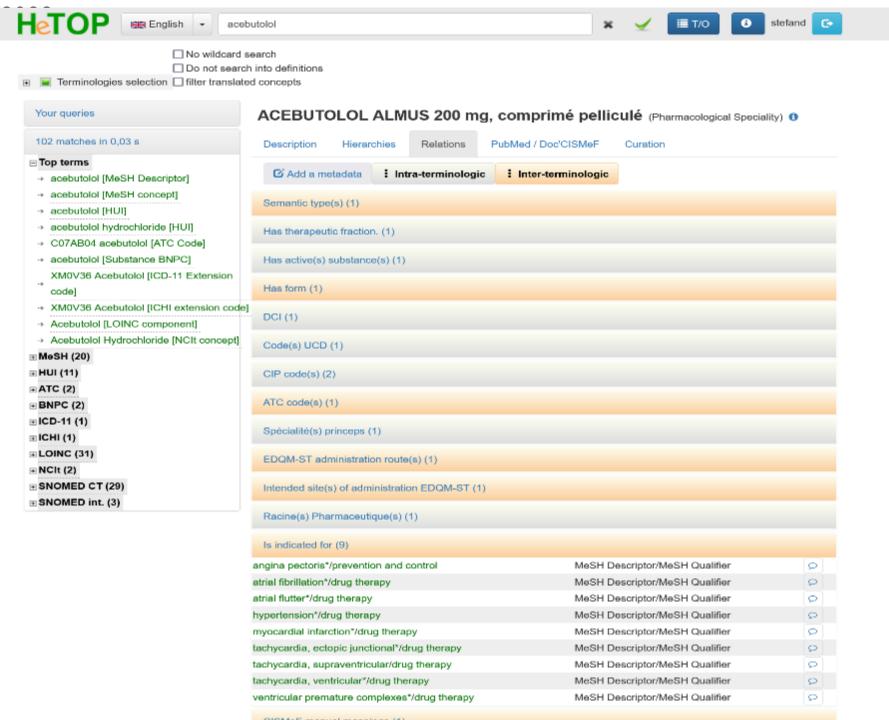
- Analytical models
- Visualization
- Statistical analysis
- etc...

### Data

#### **Data Resources**

- Databases
- Tables
- Records
- etc.





# Attributs et Objets Valeurs des attributs

- Les valeurs d'attributs sont des nombres ou symboles
- Distinction entre attributs et valeurs d'attributs
  - Un même attribut peut avoir plusieurs representations donc plusieurs valeurs différentes
    - Exemple: poinds en grammes, kilogrammes (voire pounds! ou autres)
  - Plusieurs attributs peuvent correspondre à des mêmes jeux de valeurs
    - Exemple : des identifiants uniques sont des nombres entiers



# Attributs et Objets **Types d'Attributs**

#### Nominal

Exemples : nombres, couleurs des yeux, codes postaux

#### Ordinal

Exemples: rangs (e.g., notes entre 1 et 10, hauteur {grand, moyen, petit}

#### Interval

 Exemples : dates, températures en Celsius ou Fahrenheit

#### Ratio

 Exemples : température en Kelvin, longueur, comptes, temps écoulé



# Attributs et Objets Attributs discrets et continus

#### **Attribut discret**

- Une valeur finie de valeurs d'attribut
- Exemples : codes postaux, mots de documents
- N'importe quel type de valeur (nombres, symboles, etc.)
- Note : attributs binaires très utilisés dans ce cas

#### **Attribut continu**

- Nombres réels
- Exemples : température, taille, poids
- Nombres "flottants" généralement, précision à définir (!)



# Attributs et objets **Types d'attributs**

- Le choix d'un type de donnée à utiliser pour un attribut n'est pas toujours simple
  - discret / continu, plutôt simple
  - interval, jeu de valeur, etc.
  - très important pour le stockage et l'analyse des données
  - "pas de solution miracle": demander de l'expertise à ceux qui savent ou apprenez vous même avec l'expérience



### Types de Données Caractéristiques importantes de la donnée

#### Dimension (nombre d'attributs)

Haute dimension = centaine ou milliers d'attributs (nombre de colonnes) => challenge de la big data

#### Taille

 La façon de stocker et analyser la donnée depend aussi de sa taille (nombre de lignes) => autre challenge de la big data

#### Exhaustivité

Seule la présence compte!

#### Qualité

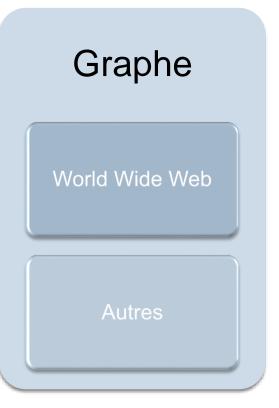
- Attention aux mélanges de valeurs, de types de valeurs pour un seul et même attributs ; et toutes autres erreurs (l'erreur est humaine)
- Particulièrement important en Santé (criticité + qualité pauvre dans la vraie vie!)

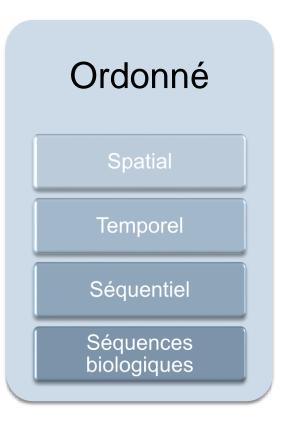


### Types de données

Différentes representations pour différents usages









# Types de données Collections d'enregistrements = tableau

 Un ensemble de métadonnées (colonnes = attributs) sous forme de tableau bidimensionnel

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



# Types données Données en matrices

Idem que précédemment mais avec plus de 2 dimensions



### Types de données Données des documents

- Exemple : Chaque document devient un "vecteur de termes"
  - Chaque terme est un élément du vecteur
  - La valeur est le nombre de fois où l'élément apparait dans chaque document

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0



### Types de données Données de transaction

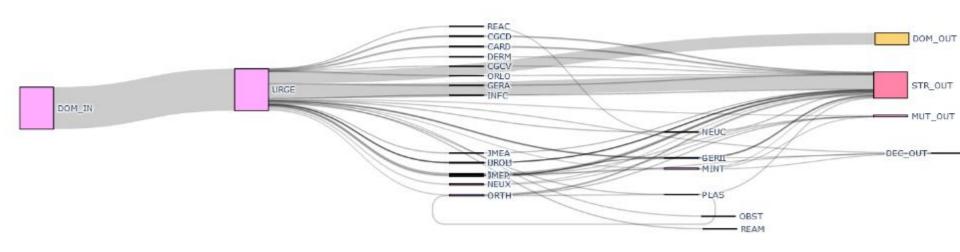
- Un type special de données, où
  - Chaque transaction implique un nombre d'éléments
  - Voici une représentation possible (achats dans un super-marché ici pour chaque client)

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk



### Types de données Données de transaction

 En santé, typiquement le parcours de soin du/des patient(s)





### Types de données Données ordonnées

#### Données de séquences biologiques

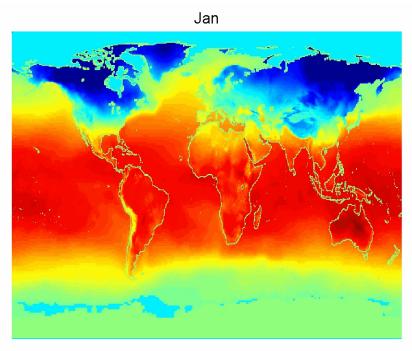
# Que ce passe-t-il si on change l'ordre?!?





#### Types de données Données ordonnées

#### Données spatiales, temporelles et spatiotemporelles

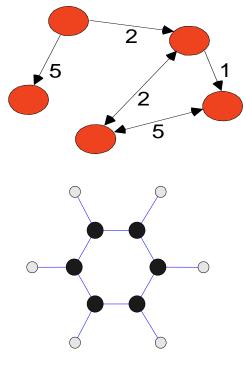






## Types de données Graphes

 Exemples: graphe générique, une molécule, site web (pages via hyperliens)



Benzène: C6H6

#### **Useful Links:**

- Bibliography
- · Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - o The Data Mine

#### Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

#### **Knowledge Discovery and Data Mining Bibliography**

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

#### **General Data Mining**

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

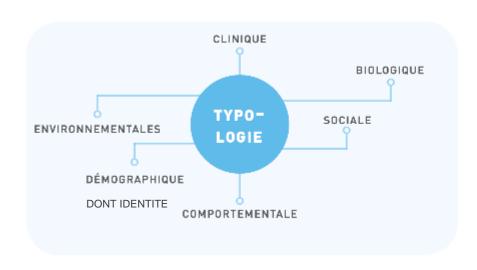
### Données de santé

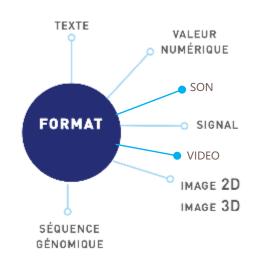
- Données qui sont « par nature » relatives à l'état de santé d'une personne (celles issues de la relation de soin par exemple)
- Mais aussi les constantes physiologiques et caractéristiques morphométriques de l'homme sain (définition de la santé!)
- Données qui le seraient compte tenu de leur destination (telles que celles issues de certains objets connectés, avec différents niveaux de granularité : exemple des trains de glycémies répétées).



## Typologie et format des données

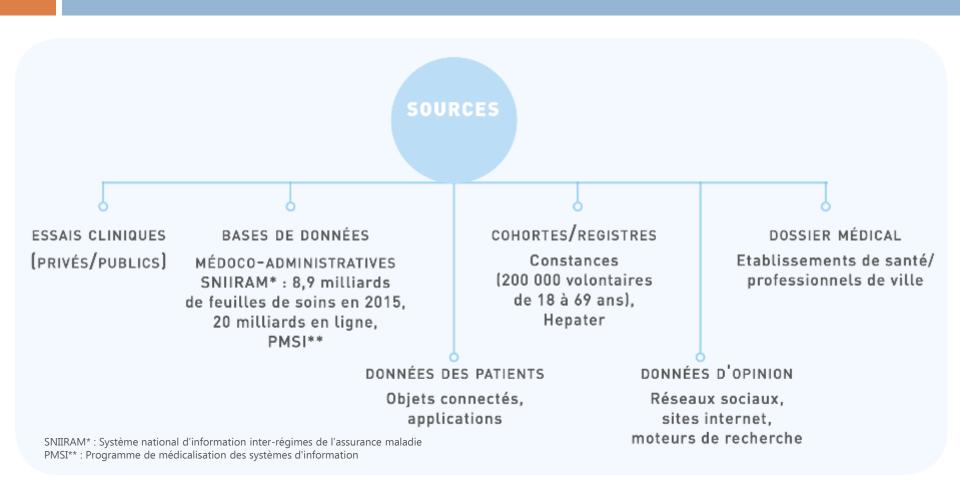
 On observe une grande disparité des données de santé que ce soit au niveau de leur typologie, de de leur format...







### Sources des données de santé





## Origine des données

- Des distinctions doivent être faites, qui portent sur l'origine des données et sur l'objectif de leur collecte.
   Il y a essentiellement trois types de sources de ce point de vue :
  - Données recueillies en vue de la gestion et du financement (en général dans des bases exhaustives)
  - Données recueillies pour la traçabilité des soins (les dossiers médicaux des établissements et des professionnels, jusqu'au Dossier médical personnel (DMP)
  - Données recueillies par enquête pour la santé publique, la veille sanitaire, la recherche épidémiologique.

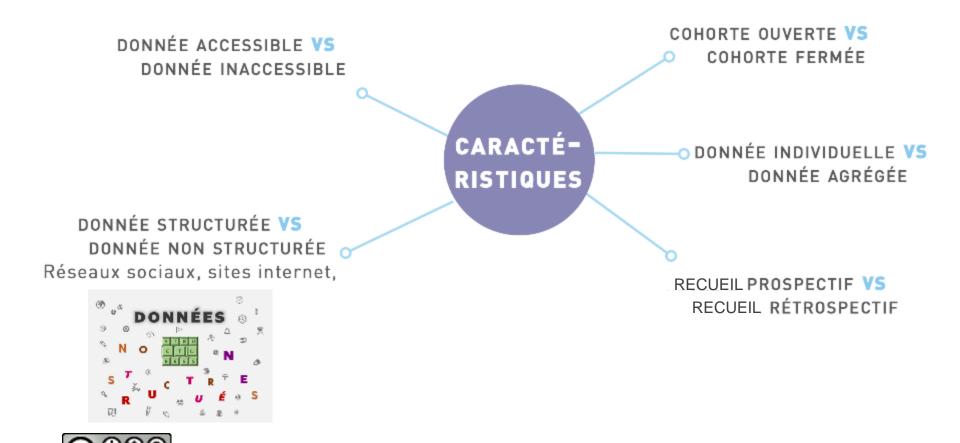


## Origine des données

- Le développement de multiples services numériques, la santé mobile, l'internet des objets génèrent un accroissement exponentiel du volume des données produites, qu'il s'agisse de plateformes de prise de rendez-vous en ligne, de systèmes pour notifier des effets indésirables, d'applications mobiles pour mieux suivre son traitement, de sites de ventes en ligne de médicaments, de réseaux sociaux développés par des communautés de patients échangeant des informations sur leur maladie, leur traitement et leur expérience (comme *PatientLikeMe* aux États-Unis, ou *Carenity* en France), etc.
- Ces sources de données débordent largement la sphère de la maladie et du soin, avec les objets connectés et des applications de bien-être ou d'activité sportive, ou avec l'exploitation des traces numériques de l'activité des individus sur internet ou dans les réseaux sociaux généralistes.
- Dans le domaine de la santé comme dans de nombreux autres domaines de l'activité humaine, la généralisation de la numérisation et l'augmentation continue des possibilités de stockage et de traitement de l'information permettent d'exploiter d'énormes volumes de données issues de sources diverses, structurées (diagnostics codés, résultats de tests, remboursements de soins…) ou non structurées (comptes rendus d'hospitalisation, échanges sur les réseaux sociaux, etc.).



### Caractéristiques des données de santé



#### Structured Data

Un-Structured Data

Medications Demographics Encounters Diagnoses Procedures Diagnostics (Ordered) Diagnostics (Results) Genetics Social History **Family History** Symptoms Lifestyle Socioeconomic

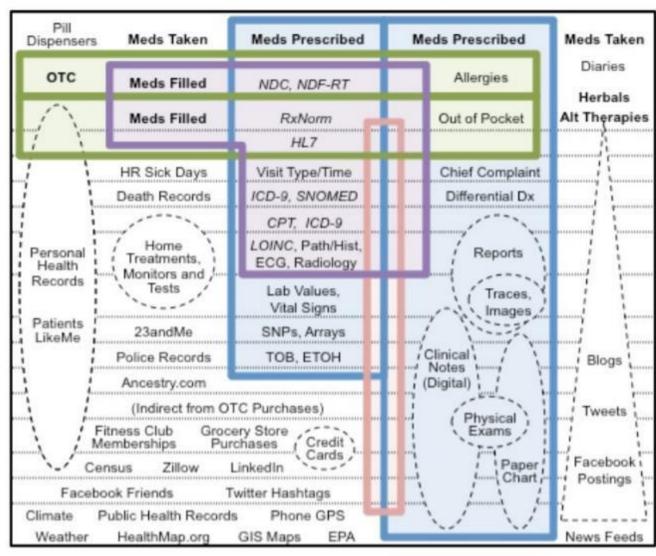
Sol H. Big Data en santé : données concernées, usages, entrepôt bio-hétérogènes et outils d'exploitation, 2016

(http://numerique.anap.fr/publication/1505-bigdata-en-sante-donnees-concernees-usagesentrepot-bio-heterogenes-et-outils-dexploitation, last visit: 23.10.17)



Social Network

Environment



Pharmacies

Claims Data, Medicare Registries, Clinical Trials Health Care Centers Data Outside the Healthcare System

### Données en vie réelle

- On désigne sous le terme « données de vie réelle », ou « données de vraie vie », des données qui sont sans intervention sur les modalités usuelles de prise en charge des malades et ne sont pas collectées dans un cadre expérimental (le cadre notamment des essais randomises contrôles, ECR), mais qui sont générées à l'occasion des soins réalisés en routine pour un patient, et qui reflètent donc a priori la pratique courante.
- Ces données peuvent provenir de multiples sources : elles peuvent être extraites des dossiers informatisés de patients, ou constituer un sous-produit des informations utilisées pour le remboursement des soins ; elles peuvent être collectées de manière spécifique, par exemple dans le cadre de procédures de pharmacovigilance, ou pour constituer des registres ou des cohortes, ou plus ponctuellement dans le cadre d'études ad hoc ; elles peuvent également provenir du web, des réseaux sociaux, des objets connectés, etc.



### Données relatives à la santé

- Les catégories de données relatives à la santé sont les suivantes :
  - Données personnelles sur les citoyens / patients. Même si des moyens techniques spécifiques sont utilisés (hébergeurs agréés pour les données partagées, numéro d'identification spécifique, carte de professionnel de santé, etc.), les données relèvent du droit commun pour la protection des données individuelles et sont donc sous le contrôle de la CNIL.
    - Les techniques sont de la responsabilité principalement de l'Agence du Numérique en santé (ANS, ancienne Agence des systèmes d'information partagée de santé – ASIP Santé), de la Caisse nationale d'assurance maladie des travailleurs salariés et du GIE SESAM-Vitale.
  - Données agrégées, statistiques épidémiologiques etc. qui résultent toujours de traitements de données individuelles collectées pour la gestion ou pour des enquêtes et études spéciales.
  - Données sur l'offre caractéristiques et activité des hôpitaux, tarifs de professionnels etc. Celles-ci approchent une autre problématique, fréquente pour les données publiques : la protection de l'information sur l'entreprise.



## Qualité des données

Deux visions
Celle de l'ingénieur et
Celle du professionnel de santé

### Qualité des données

- La qualité des données repose sur leur fiabilité.
- Lorsqu'on dispose de données actualisées de grande qualité, on peut les utiliser en toute confiance dans le cadre
  - Du processus de soins (traçabilité, suivi et continuité des soins, démarche décisionnelle)
  - De travaux de recherches utiles,
  - De planification stratégique pertinente,
  - De gestion de la délivrance des soins de santé.



#### Définition

- Les données de santé sont régies par la loi informatique et libertés (la loi n°78-17 du 6 janvier 1978) définissant les "données à caractère personnel" et la jurisprudence.
- Le nouveau Règlement Général européen sur la Protection des Données (RGPD ou GPDR, acronyme anglais), applicable depuis mai 2018, poursuit comme objectifs de renforcer les droits des personnes et de responsabiliser les acteurs autour des données, et en particulier les données de santé, qui ont leur définition propre, soit des "données à caractère personnel relatives à la santé physique ou mentale d'une personne physique, y compris la prestation de service de soins de santé, qui révèle des informations sur l'état de santé de cette personne".



### Enjeux de la qualité des données





### Les critères de qualité

- la pertinence et la complétude
- la fiabilité
- la validité, la cohérence
- l'exactitude, la précision
- l'actualité, la régularité
- la compréhension, l'intelligibilité
- l'accessibilité



#### Pertinence

- Les données répondent à la question posée.
  - Par exemple, pour tirer des conclusions d'une étude sur l'efficacité d'une prophylaxie, il est nécessaire de disposer de données sur la date de début du traitement, son intensité, la proportion des doses prescrites réellement administrées, l'évaluation de la fonction clinique concernée, ou les résultats d'une évaluation de données biologiques ou d'un score de qualité de vie...
- La pertinence d'une enquête, d'une étude, d'un dispositif, est son utilité; elle dépend de la connaissance et de la maîtrise du domaine du prestataire ou propriétaire des données, et des bonnes connaissances et applications des traitements et usages.



### Complétude

- On est souvent confronté à cette problématique des valeurs manquantes qui rend les données incomplètes.
- On peut atténuer cette problématique en s'assurant d'utiliser une bonne source de données.



#### Fiabilité

- Dans la mesure du possible, les données correspondent à la situation réelle; dans certains cas, l'approximation, mais pas l'hypothèse, peut être acceptable.
- L'impact des différents milieux et traitements ou des changements au fil du temps ne peut être mesuré qu'en utilisant des données précises.
  - Par exemple, si un registre de patients ne prend pas en compte les cas de décès ou d'émigration, cela pourrait conduire à une surestimation de la population de patients réelle et à des conclusions erronées sur la quantité de facteurs utilisée par patient.



#### Validité

- Degré de conformité des données aux règles ou contraintes définies.
  - Les types de données : les valeurs d'une colonne doivent être d'un type de données particulier, par exemple, numérique, date, etc.
  - Contraintes de plage : par exemple, les nombres doivent être compris dans une plage donnée.
  - Contraintes obligatoires : par exemple certaines colonnes ne peuvent pas être vide.
  - Unicité : un champ ou plusieurs champs combinés doit être unique dans un dataset.
  - Clé étrangère : comme pour les bases de données relationnelles, la colonne de clé étrangère ne peut pas avoir une valeur qui n'existe pas dans la clé primaire référencée.
  - Motifs d'expression régulière : concernent des champs de textes doivent respecte un format précis. Exemple les numéros de téléphone qui doivent respecte le format (+33) 6 66 66 66.
  - Validation entre champs : concernent des conditions qui doivent être remplies. Par exemple, une date de décès ne pas être avant une date de naissance de la même personne.



#### Cohérence

La cohérence consiste à la validation interne de la base de données, mais aussi et surtout à la comparabilité des données et des résultats à des connaissances antérieures, en particulier si le dispositif est répété dans le temps, comme c'est le cas des panels.



#### Exactitude et Précision

- Différence entre exactitude et précision.
  - Par exemple, dire qu'on vit en Europe est vrai. Cependant, cette réponse n'est pas précise.
- Ce qu'on doit vérifier est la précision des données et pas seulement leur exactitude.
- Cette tache n'est clairement pas simple. Car définir toutes les valeurs valides possibles permet de repérer facilement les valeurs non valides, cela ne signifie pas pour autant qu'elles sont exactes et encore moins qu'elles sont précises.



#### Actualité

- Appelée aussi récence
- C'est le temps qui s'écoule entre la collecte des données et la parution des résultats.
- De façon plus générale, ce terme est adapté au temps qui s'écoule entre le moment observé et le moment du recueil lui-même.
- Intérêt d'horodater la donnée saisie mais aussi d'horodater le contenu de la donnée recueillie
  - Le JJ/MM/AAAA, la donnée a été saisie
  - La donnée correspond à un événement en date du JJ/MM/AAAA



# Régularité

 Les données doivent être collectées avec rapidité (sondages uniques) ou une fréquence (collecte régulière de données, telle que les registres) convenant à l'usage prévu.



# Compréhension - Intelligibilité

- Les données doivent être collectées en utilisant la terminologie technique standard.
- Il est également important de les retranscrire dans un langage que les utilisateurs ciblés comprennent.
- L'intelligibilité porte sur la documentation de la méthodologie employée, sa clarté, sa compréhension par des utilisateurs non-spécialistes.
- La lisibilité (readability) est la facilité avec laquelle un lecteur peut comprendre un texte écrit. En langage naturel, la lisibilité d'un texte dépend de son contenu et de sa présentation. Les chercheurs ont utilisé divers facteurs pour mesurer la lisibilité, tels que : Vitesse de perception Perceptibilité à distance



#### Accessibilité

- L'accessibilité concerne le mode de restitution, de mise à disposition, de présentation.
- La visualisation en fait partie.



### Collecter des données de qualité

- Définir clairement la question à laquelle il faudra répondre et s'assurer que les données collectées sont pertinentes.
- Utiliser des outils valides pour collecter les données.
  - Les outils validés sont des méthodes de collecte de données qui ont été évaluées et jugées fiables.
- Essayez de mesurer des aspects ou événements simples, objectifs et quantifiables plutôt que complexes ou subjectifs
  - par exemple, calculer le taux d'absentéisme à l'école plutôt que le niveau d'études.
- Les méthodes de mesure doivent être reproductibles en vue de donner des résultats similaires si vous ou une autre personne les utilisiez à nouveau afin de mesurer le même phénomène.



### Collecter les données médicales

- Via les professionnels de santé et la R&D. Les données peuvent être récupérées directement par les professionnels de santé (dossiers médicaux des patients lors d'une hospitalisation ou d'une visite chez le médecin) ou bien grâce à des examens plus poussés (lors d'essais cliniques ou grâce à l'analyse de l'ADN). Le volume de données exposé est donc très important.
- Via les individus eux-mêmes. Les individus produisent euxmêmes leurs données au moyen d'objets connectés (montres, vêtements, domotique, applications mobiles). Ces éléments constituent une source précieuse d'informations en nous renseignant sur l'activité physique, le rythme cardiaque, les heures de sommeil et place l'individu comme protagoniste voire acteur principal de sa propre santé.



# « Open data » ou données ouvertes

### Données ouvertes

- L'ouverture des données d'intérêt public vise à encourager la réutilisation des données au-delà de leur utilisation première par l'administration.
- En utilisant, directement ou via des applications, des données publiées sur la plateforme <u>data.gouv.fr</u>, on peut par exemple :
  - répondre à des questions ;
  - prendre des décisions, pour soi, sa commune, son association ou son entreprise;
  - bénéficier de services utiles au quotidien : pour se déplacer, éviter le gaspillage alimentaire, connaître les services publics à proximité de son domicile ;
  - encourager la transparence démocratique des institutions et des élus, par exemple : connaître l'utilisation de la réserve parlementaire, les budgets de l'État et des collectivités, les titres de presse aidés par l'État.





### Mentions légales

- L'ensemble de ce document relève des législations française et internationale sur le droit d'auteur et la propriété intellectuelle.
- Tous les droits de reproduction de tout ou partie sont réservés pour les textes ainsi que pour l'ensemble des documents iconographiques, photographiques, vidéos et sonores.
- Ce document est interdit à la vente ou à la location par un tiers autre que l'Université Cote d'Azur.
- La diffusion, la duplication, la mise à disposition du public (sous quelque forme ou support que ce soit), la mise en réseau, de tout ou partie de ce document, sont strictement réservées à l'Université Cote d'Azur.
- L'utilisation de ce document est strictement réservée à l'usage privé des étudiants inscrits aux cours et au tutorat organisés par l'UFR de Médecine de l'Université Cote d'Azur et de l'Université de Rouen Normandie (projet SaNuRN), et non destinée à toute autre utilisation privée ou collective, gratuite ou payante.

