Données de santé et qualité des données

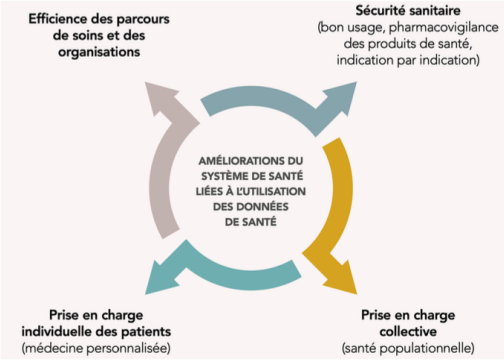
I- Position du problème

A) 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.

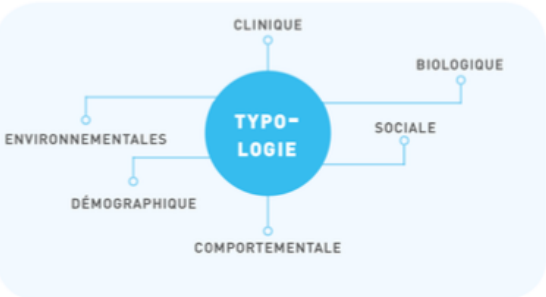
B) Enjeux des usages de la données

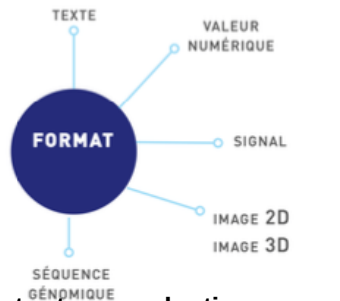


II- Données de santé

A) Définition

Les données de santé sont régies par la loi informatique et libertés (loi IFL du 6 janvier 1978) définissant les "données à caractère personnel" et la jurisprudence.
Le nouveau règlement Européen sur la Protection des Données, 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**".
Les données de santé sont « par nature » **relatives à l'état de santé d'une personne** (celles issues de la relation de soin par exemple). Mais elles sont aussi relatives aux constantes physiologiques et caractéristiques morphométriques de l'homme sain (définition de la santé).
On observe une **grande disparité des données de santé,** que ce soit au niveau de leur **typologie**, de leur **format**, etc.





B) Sources et origine des données de santé

Il ne faut pas confondre « **sources** » de données et « **producteurs** » de données +++ :
- Les sources sont le **cadre opérationnel de la production**. Elles décrivent plus l'environnement de stockage et de mise à disposition pour différents usages. On trouve dans les sources de données : des données **individuelles**, **collectives** ou **agrégées**.
- Le producteur produit la **donnée** : c'est un professionnel de santé, c'est aussi possiblement une machine.

Les données diffèrent selon leur **nature** et, en fait, leur **producteur** – d'autant que chaque acteur les développe
pour ses propres objectifs du moment.
D'autres distinctions doivent être faites, qui portent sur **l'origine des données** et sur **l'objectif de leur collecte**.
Il y a essentiellement 3 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 le **soin** (les dossiers médicaux des établissements et des professionnels, jusqu'au Dossier médical personnel)
- Données recueillies par enquête pour la **santé publique**, la **veille sanitaire**, la **recherche épidémiologique**.

Le développement de multiples services numériques, la santé mobile, l'internet des objets génèrent un accro**issement exponentiel du volume des données produites, qu'i**l 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 sporti**ve, **ou avec l'exploitati**on des traces numériques de l'**activité des indi**vidus sur internet ou dans les réseaux sociaux généra**listes.**
Dans le domaine de la santé comme dans de nombreux autres domaines de l'activité humaine, la généralisation de la numérisa**tion et l'augmen**t**ation continue des possibilités de stockage et de traitement de l'information permettent d'exploiter d'énormes volumes de données issues de sources diverses, struct**urées (diagnostics codés, résultats de tests, remboursements de soins...) ou non structurées (comptes rendus d'hospitalisation, échanges sur les réseaux sociaux, etc.).

C) Caractéristiques des données de santé

DONNÉES STRUCTURÉES OU NON STRUCTURÉES :

Données structurées = informations (mots, signes, chiffres...) contr**ôlées par de**s *référentiels et présentée*s dans des cases (les champs d'une base de données) qui permettent leur interprétation et leur traitement par des machines.
Données non structurées = le reste, tout ce qui n'est pas organisé en base de données**, c'e**st-à-dire la bureautique, la messagerie, les images, les *vidéos, etc.*
*En informatique, les informations structurées sont des info*rmations qui figurent **dans les bases de données** et les langages informatiques. On reconnaît les informations structurées au fait qu'elles sont disposées de façon à être traitées automatiquement et efficacement par un logiciel, mais pas nécessairement par un humain.
DONNÉES ACCESSIBLES - DONNÉES NON STRUCTUREES :
La France s'est doté la première d'un régime juridique spécifique aux données personnelles et à l'utilisation des données personnelles. En effet, la loi dite Informatique et Liberté promulguée le 06 janvier 1978 a pour objet spécifi**que de protéger le traitement** des données à c**aractère person**nel.

Le caractère sensible de cette catégorie de données, qui permet ainsi de catégoriser les individus en fonction de leur ethnie, sexe, état de santé, etc...., justifie à lui seul la mise en place d'une protection.
Si cette loi s'attache à traiter de la protection de l'ensemble des données dites à caractère personnel, la loi dite "**Kouchner**" promulguée le **4 mars 2002** a pour objet de s'intéresser particulièrement aux données médicales.
Ainsi, l'article L. 1111-7 du code de la santé publique met en place pour les patients les **conditions d'accès à leurs données relatives à leur santé.**
Lorsqu'un individu souhaite avoir accès à n'importe quel document dont le contenu est relatif à son état de santé (par exemple une feuille de consultation ou une ordonnance médicale), ce dernier peut demander directement ou par le biais d'un médecin l'accès à ce document.
Il est une fois de plus intéressant de noter que la loi rappelle le **caractère à la fois informatique ou non d'une donnée dite médicale** : l'interprétation de la loi ne peut ainsi pas permettre d'abus de langage allant dans le sens d'une stricte interprétation informatique du terme de "donnée".

D) 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, 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.

E) 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 (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.

III- 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.

Enjeux de la qualité des données :

A) CRITÈRES QUALITÉ :

Critères qualité = pertinence et complétude, fiabilité, validité et cohérence, exactitude et précision, actualité et régularité, compréhension et intelligibilité, accessibilité

1° 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.

2° COMPLÉTUDE : On est souvent confronté à ce problème des **valeurs manquantes** qui rend les **données incomplètes.** On peut atténuer ce problème en s'assurant d'utiliser une **bonne source de données**.

3° 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.

4° VALIDITÉ : La validité correspond au **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.
[Ex : les numéros de téléphone qui doivent respecte le format (+33) 6 66 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.

5° 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.

6° 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 tâche n'est clairement pas simple. En effet, 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.

7° ACTUALITÉ : Elle est 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 xx/xx/xxxx la donnée a été saisie, la donnée correspond à un événement en date du xx/xx/xxxx.

8° 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**.

9° COMPRÉHENSION ET 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.

10° ACCESSIBIITÉ : L'accessibilité concerne le mode de restitution, de mise à disposition, de présentation. La visualisation en fait partie.

B) 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**.
Essayer de mesurer des aspects ou évènements simples, objectifs et quantifiables plutôt que complexes ou subjectifs. [EX : 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 eux-mêmes 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 principal de sa propre santé.**

IV- 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 data.gouv.fr, 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.

Un exemple d'utilisation de l'open data : prévalence des IVG chez les adolescentes. Focus sur les Alpes-Maritimes. Une telle donnée peut ête utile pour identifier des disparités géographiques au sein d'un territoire de santé, identifier des zones où il faut engager une prévention « contraception », suivre la mise en place des mesures préventives prises, etc.

J'espère que ce cours vous aura plus. C'est le dernier qu'il a abordé lors de son premier crénau de 4H