



Université de Strasbourg

FACULTÉ DE PHARMACIE

N° d'ordre : _____

MÉMOIRE DE DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

INTÉRÊT DE LA STRUCTURATION DE LA DONNÉE EN SANTÉ

Présenté

Par **Erwan LEFEBVRE**

Soutenu le **21 septembre 2023** devant le jury constitué de

M. le Professeur Jean-Yves PABST, Président

M. le Docteur Julien GODET, MCU-PH, Directeur de thèse

M. le Docteur Loïc BROTONS, Membre du jury

Approuvé par le Doyen et
par le Président de l'Université de Strasbourg



Doyen	Ether KELLENBERGER
Directeurs adjoints	Julien GODET
	Béatrice HEUTAUT
	Emilie SICK
Directeur adjoint étudiant	Léo FEBRENA-MOURIAUX

LISTE DU PERSONNEL ENSEIGNANT

Professeurs :

Philippe	BOUCHER	Physiologie
Nathalie	BOULANGER	Stressologie
Lise	BOUREL	Chimie thérapeutique
Pascal	DIGRE	Stomatologie
Said	ESNAHAR	Chimie analytique
Valérie	GEOFFROY	Microbiologie
Philippe	GEBREL	Bactériologie, Virologie
Jean-Pierre	GIES	Pharmacologie moléculaire
Béatrice	HEUTAUT	Pharmacie galénique
Esther	KELLENBERGER	Bio-informatique
Maxime	LEHMANN	Biologie cellulaire
Eric	MADCHIN	Chimie analytique
François	MEGERLIN	Droit et économie pharm.
Yves	MELY	Physique et Biophysique
Jean-Yves	MABET	Droit Economie pharm.
Françoise	ROSS	Toxicologie
Valérie	SCHNE-REITH	Pharmacologie
Florence	TOTI	Pharmacologie
Thierry	VANDANNE	Biogalénique
Catherine	VONTHRON	Pharmacognosie
Pascal	WEHLE	Pharmacie galénique

Professeurs-praticiens hospitaliers

Julien	GODET	Biophysique - science des données
Jean-Marc	LESSINGER	Biochimie
Benoit	MICHEL	Pharm. clinique santé publique
Pauline	SOULAS-SPRAUEL	Immunologie
Geneviève	UREAUD-SÉQUINÉ	Pharmacocinétique

Enseignants contractuels

Alexandra	CHAMPEL	Pharmacie d'officine
Matthieu	FOHRE	Pharmacie d'officine
Philippe	GALAS	Droit et économie pharm.
Philippe	NANDE	Ingénierie pharmaceutique
Caroline	WILLET - WEHLE	Pharmacie d'officine

Maîtres de Conférences :

Nicolas	ARTON	Pharmacie biogalénique
Farouha	BATTOU	Biochimie
Martine	BEGAENTZLE	Chimie analytique
Elisa	BOMBARDA	Biophysique
Aurélien	BOUBERDUX	Pharmacochimie
Emmanuel	BOUTANT	Virologie et Microbiologie
Virginique	BRUBAN	Physiologie et physiopath.
Anne	CASSET	Toxicologie
Thierry	CHATRENEAU	Pharmacologie
Manuela	CHEREI	Pharmacie biogalénique
Guillaume	CONZATTI	Pharmacie galénique
Marcella	DE GIOSSI	Pharmacochimie
Serge	DAIMONT	Biologie cellulaire
Gisèle	HAAN-ARCHIPOFF	Plantes médicinales
Cédric	JACQUEMARD	Chémothérapie
Julie	KARRENKO	Pharmacochimie
Sonia	LORDEL	Chimie analytique
Clarissa	MAEHLING	Chimie physique
Bachir	MATZ-WESTPHAL	Pharmacologie
Charifa	MEHADI	Chimie
Nathalie	NEIDERHOFER	Pharmacologie
Sergie	ORTIZ AGUIRRE	Pharmacognosie
Sylvie	PERGOTEY	Parasitologie
Bumain	PERTSCH	Chimie en flux
Frédéric	PRZYBYLLA	Biostatistiques
Patrice	RASSAM	Microbiologie
Eléonore	REAL	Biochimie
Andreas	REISCH	Biophysique
Ludvine	RIFFAULT-VALDES	Analyse du médicament
Carole	RONZANI	Toxicologie
Emilie	SICK	Pharmacologie
Yaouba	SOUABOU	Pharmacognosie
Maria-Vittoria	SPANEIDA	Chimie thérapeutique
Jérôme	TERRIARD	Physiopathologie
Bassara	TOLWS	Chimie physique
Aurélien	UREAUD	Pharmacognosie
Bruno	VAN OVERLOOP	Physiologie
Maria	ZENOU	Chimio-génomique

Maîtres de conférences - praticiens hospitaliers

Julie	BRUNET	Parasitologie
Nelly	ETIENNE-SELLOUM	Pharmacologie- pharm. clinique

Assistant hospitalier universitaire

Damien	REITA	Biochimie
--------	-------	-----------

SERMENT DE GALIEN

JE JURE,

en présence des Maîtres de la Faculté,
des Conseillers de l'Ordre des Pharmaciens
et de mes condisciples :

D'honorer ceux qui m'ont instruit
dans les préceptes de mon art et de
leur témoigner ma reconnaissance en
restant fidèle à leur enseignement ;

D'exercer, dans l'intérêt de la santé publique,
ma profession avec conscience et de respecter non
seulement la législation en vigueur, mais aussi les règles
de l'honneur, de la probité et du désintéressement ;

De ne dévoiler à personne les secrets
qui m'auront été confiés et dont j'aurai eu
connaissance dans la pratique de mon art.

Si j'observe scrupuleusement ce serment,
que je sois moi-même honoré
et estimé de mes confrères
et de mes patients.

Table des matières

FICHE SIGNALÉTIQUE.....	7
I. Introduction :.....	8
1. Contexte et justification du sujet	8
2. Objectifs de la thèse	9
II. État de l'art	10
1. Qu'est-ce que la donnée structurée.....	10
2. Les modèles de bases de données :	14
3. La donnée de santé	16
4. Les niveaux de structuration	17
5. Revue des directives pour structurer, acquérir, extraire et normaliser la donnée.....	21
3. Exemples de nomenclatures	24
a. Classification Internationale des Maladies (CIM-11 ou ICD-11) :.....	24
b. SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) :.....	26
4. Le DPI.....	28
5. Mon Espace Santé et le DMP	30
7. Règlementation en vigueur en France (HDS, RGPD)	34
8. Entrepôts de santé	37
9. Règlementation aux États-Unis	38
III. Application.....	40
1. Identification des lacunes du système de santé (cybersécurité, perte de données, structure, interopérabilité)	40
2. Application au domaine de la santé	44
IV. Perspectives	47
1. Revue des projets existants et futurs sur la donnée structurée en santé.....	47
a. Le SNDS et le HDH	49
b. Les données structurées dans la pandémie :	50
2. L'intérêt pour la recherche médicale	51
3. Intérêt pour les algorithmes d'IA (Machine learning, deep learning)	53
4. Défis des données structurées	55
5. Intérêt pour les essais in silico.....	56
6. Intérêt pour les jumeaux numériques.....	57
V. Conclusion	58
Résumé des résultats de la thèse et perspectives	58
VI. Références.....	59
Liste complète des références utilisées dans la thèse	59

Liste d'abréviations utilisées :

- ANS : Agence Nationale de la Sécurité
- API : Application Programming Interface
- CCAM : Classification Commune des Actes Médicaux
- CépiDC : Centre d'épidémiologie sur les causes médicales de Décès
- CIM-11 : Classification Internationale des Maladies, 11e révision
- CNAM : Caisse Nationale d'Assurance Maladie
- CNIL : Commission Nationale de l'Informatique et des Libertés
- CPS : Carte de Professionnel de Santé
- DPI : Dossier Patient Informatisé
- DP : Dossier Pharmaceutique
- DPO : Data Protection Officer
- DSI : Direction des Systèmes d'Information
- ECG : Électrocardiogramme
- EDHS : European Health Data Space
- EHR : Electronic Health Record
- ETP : Équivalents Temps Plein
- IRM : Imagerie par Résonance Magnétique
- MCO : Médecine, Chirurgie, Obstétrique
- NABM : Nomenclature des Actes de Biologie Médicale
- NGAP : Nomenclature Générale des Actes Professionnels
- NLP : Natural Language Processing
- OMS : Organisation Mondiale de la Santé
- OTC : Over-The-Counter
- PDS : Permanence Des Soins
- PMI : Protection Maternelle et Infantile
- PMSS : Plafond Mensuel de la Sécurité Sociale
- RGPD : Règlement Général sur la Protection des Données
- RSSI : Responsable de la Sécurité des Systèmes d'Information
- SAAS : Software as a Service
- SIH : Système d'Information Hospitalier
- SEO : Search Engine Optimization
- SNIIRAM : Système National d'Information Inter-Régimes de l'Assurance Maladie
- SNOMED-CT : Systematized Nomenclature of Medicine-Clinical Terms
- SQL : Structured Query Language

- T2A : Tarification à l'Activité

- W3C : World Wide Web Consortium

Remerciements :

C'est la fin d'un cycle pour moi et je me dois de remercier tous ceux qui m'ont soutenu dans ce projet.

Merci tout d'abord au professeur Jean-Yves PABST de diriger cette thèse, ainsi qu'au docteur Julien GODET, jury, pour son soutien dès le départ et son accompagnement.

Je tiens à remercier chaleureusement toute l'équipe de GALEON et notamment Dr. Loïc BROTONS pour avoir cru en moi et donné une chance de prouver ma valeur en tant que pharmacien junior. Son mentorat et sa patience sont des valeurs qui m'ont touché durablement.

Merci à Nicolas PAGES et Nicolas DELRUE d'être les meilleurs collègues que l'on puisse rêver.

Merci à tous les autres pour cette ambiance et ce beau projet commun.

Merci papa et maman de m'avoir forcé à continuer mes études plutôt que la musique, je serai un pharmacien musicien grâce à vous. À ma petite sœur qui trouvera sa voie également j'en suis sûr.

Je tenais à remercier l'équipe enseignante et notamment Catherine LEDIG pour son implication sans faille et sa disponibilité.

Merci à Chloé qui m'apporte bien plus qu'elle ne le pense.

Merci à Julien et à Thomas pour leur amitié qui dure malgré les années.

Merci enfin à la Tibetan Fox pour toutes ces années de rigolade et d'entraide, et à d'autres encore.

FICHE SIGNALÉTIQUE

LEFEBVRE Erwan

Né le 19/07/1999 à Mulhouse

INTÉRÊT DE LA DONNÉE STRUCTURÉE EN SANTÉ

Soutenu le 21 septembre 2023 à Illkirch-Graffenstaden

N° d'ordre : _____

RÉSUMÉ

Dans cette thèse, nous explorons l'importance de la structuration des données dans le domaine de la santé. Nous commençons par analyser les avantages de la donnée structurée, notamment en termes d'interopérabilité et de qualité des données, et identifions également les défis associés, tels que la nécessité d'un changement culturel et organisationnel pour les professionnels de santé. Nous discutons de comment la donnée structurée touche divers aspects du métier de pharmacien, que ce soit en milieu hospitalier, en officine ou dans l'industrie pharmaceutique. Nous mettons en évidence l'impact des technologies de l'information sur la transformation du métier et comment elles peuvent améliorer l'arsenal thérapeutique. Nous abordons également l'avènement des technologies d'intelligence artificielle dans la recherche médicale. Nous soulignons les avancées de la France en matière de données structurées et montrons que des mesures concrètes sont prises pour avancer dans la bonne direction. En plus de ces aspects techniques, nous examinons les implications réglementaires et organisationnelles de la donnée structurée en santé. Nous mettons également en lumière l'importance de la collaboration et de l'open source dans ce domaine. Nous montrons comment la maîtrise des technologies peut servir notre système de santé, non seulement en améliorant les soins aux patients mais aussi en alimentant la recherche. Une perspective sur les projets existants et futurs sur la donnée structurée en santé, y compris l'intérêt pour la recherche médicale, les algorithmes d'IA, les défis des données structurées, et l'intérêt pour les essais in silico et les jumeaux numériques.

Mots clés : Données – Structuration – Politique – Droit – IA – Interopérabilité – Open Source

Directeur de Thèse : Jean-Yves PABST

I. Introduction :

1. Contexte et justification du sujet

Ces dernières années ont vu une fascinante accélération des avancées dans le domaine de l'informatique, de même qu'un accroissement majeur de la quantité de données produites par tous les secteurs, dont celui de la santé.¹ Ces volumes parfois indomptables de bytes gagnent à être organisés, normalisés, structurés dans des standards et former, dès la saisie, de la donnée structurée. La COVID-19 a accéléré la numérisation d'un univers déjà très tourné vers la technologie et nombre d'initiatives institutionnelles vont en ce sens. La doctrine française en termes de données de santé gagne en cohérence et l'apparition de plateformes comme Mon Espace Santé, ou d'initiatives nationales comme le Health Data Hub sont le fruit d'une réflexion anticipée et profuse en la matière. Ces transformations progressent dans le sens du concept de l'Open Data, qui vise à rendre les données collectées par les autorités publiques accessibles à la société civile. On observe également et plus que jamais, l'entrée du Big Data en santé.² La donnée est plus que jamais une ressource phare de l'innovation et de plus en plus d'acteurs prennent conscience de sa valeur, notamment dans le milieu de la santé. Cela peut être le motif d'actes malhonnêtes pour tenter de prendre le contrôle ou revendre ces mêmes données, comme les cas toujours plus nombreux de cyberattaques sur les centres de soins en sont témoin.³

Pour autant, produire et stocker des quantités impressionnantes de données n'est pas la finalité. En effet, on observe une sous-exploitation notable et un manque de valorisation médicale et financière de la part des acteurs publics et privés. La gestion, l'analyse et l'exploitation efficaces de ces données représentent un enjeu majeur pour améliorer la qualité des soins, optimiser les ressources et favoriser la recherche médicale.

La donnée structurée présente un fort potentiel pour transformer le secteur de la santé. L'exploitation de ce potentiel, bien que primordiale, implique des changements majeurs à tous les niveaux de l'industrie de la santé, chaque changement pouvant lui-même constituer le sujet d'une thèse séparée. Toutefois, l'utilisation optimale de la donnée structurée

¹ <https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde/>

² <https://www.inserm.fr/dossier/big-data-en-sante/>

³ https://umap.openstreetmap.fr/fr/map/attaques-cybersecurite-aupres-dorganismes-publics_821557#7/47.610/5.757

s'accompagne d'autant de défis pour le soignant comme pour le technicien ou le juriste. Implémenter ces technologies implique l'acquisition de matériel nécessitant une infrastructure technique robuste et une compliance réglementaire adéquate pour assurer la disponibilité, la confidentialité et la sécurité des données, en accord avec les réglementations locales. Il est donc essentiel de conjuguer les compétences de chaque domaine concerné par la donnée structurée. De plus, la transition vers des systèmes de données structurées implique un changement culturel et organisationnel pour les professionnels de santé, qui doivent s'adapter à de nouveaux outils, processus et quelquefois, responsabilités. Tout système de santé parvenant à tirer profit de ces mécanismes de structuration pourrait en sortir avec un avantage considérable, notamment par la quantité et la qualité de données qu'il produirait pour alimenter la recherche.

2. Objectifs de la thèse

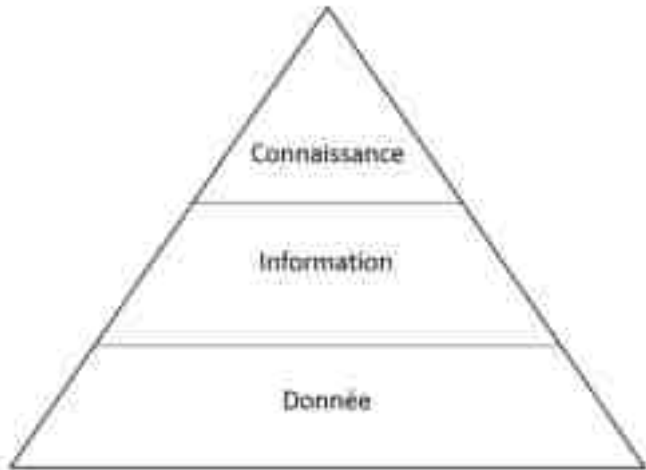
La présente thèse se propose d'étudier l'intérêt de la donnée structurée en santé, en analysant ses avantages et ses défis, ainsi que les conditions requises pour sa mise en œuvre et son utilisation optimale. J'ai choisi ce sujet car il concerne tout autant le pharmacien hospitalier avec ses logiciels intégrés au système d'information de l'hôpital (SIH) que celui d'officine et sa base de données de médicaments, tout comme le pharmacien d'industrie qui va utiliser ces données pour aider à améliorer l'arsenal thérapeutique par exemple. J'ai également une appétence toute particulière pour le croisement des technologies de l'information et du changement qu'elles opèrent dans la transformation du métier de pharmacien en lui-même. Il en est venu une appellation qui s'apparentait au départ à une plaisanterie, mais qui est aujourd'hui devenu mon métier et mon titre officiel au sein de mon entreprise, celui de cyberpharmacien. J'ai voulu à travers cette thèse parler de ma pratique au quotidien et des aspects méconnus du système de santé à la croisée entre compétences techniques et protection de la donnée personnelle. De plus, les innovations techniques citées dans cette thèse ont souvent un lien avec un projet collaboratif voire open source, ce qui est un mode de fonctionnement qui me plaît particulièrement et que je souhaite mettre en lumière. Ce sujet me paraissait également montrer à quel point maîtriser l'avènement des technologies d'intelligence artificielle peut servir la recherche et notre système de santé, au travers de techniques comme le Natural Language Processing (NLP) ou le machine learning. Je voulais également souligner les avancées de la France en la matière et montrer que des

mesures concrètes sont prises pour avancer dans la bonne direction. En abordant les aspects techniques, organisationnels et réglementaires, cette recherche vise à contribuer à une meilleure compréhension des enjeux liés à la donnée structurée en santé et à fournir des recommandations pour faciliter son adoption et son exploitation au service de la qualité des soins et de la recherche médicale. Il est selon moi très important qu'un sujet, d'apparence très technique, soit traité autant d'un point de vue des soignants que par des techniciens.

II. État de l'art

1. Qu'est-ce que la donnée structurée

Une donnée est une valeur ou un fait brut qui n'a pas encore été interprété. Elle peut être matérialisée de manière numérique ou textuelle à travers des symboles, des images, des tableaux, des représentations graphiques. Elle s'oppose à l'information, que l'on peut traduire dans un certain contexte par intelligence en anglais. C'est une interprétation organisée de la donnée pour en tirer une conclusion, un sens. C'est à ce concept qu'on se réfère lorsque l'on parle d'intelligence artificielle, par exemple. Une information correctement digérée et rangée peut devenir à terme de la connaissance, véritable statut final de la donnée et qui fait avancer un art, une science ou une technique dans sa globalité. Mais nos esprits bien organiques, eux, ont compris que la donnée collectée sans structure, diffuse, n'est pas la plus utile. Il peut être nécessaire de lui apposer une architecture précise prédéfinie en amont, avant même sa collecte, afin de faciliter son analyse, son traitement, ainsi que son stockage. C'est ce que l'on nomme le processus 'du schéma à l'écriture'. Les données sont alors formatées dans des champs précis, normés, qui peuvent facilement faire l'objet d'une requête Structured Query Language (SQL), par exemple.



Certaines données servent parfois à en décrire d'autres pour mieux les définir, il s'agit des métadonnées. On peut penser aux coordonnées géographiques incluses dans les propriétés d'un fichier image pris par un téléphone par exemple. Les métadonnées sont des vecteurs essentiels de l'interopérabilité, qui permettent aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs. **DICOM dans la santé (radios)** Ce partage est par exemple rendu possible sur Internet par le Web sémantique, défini par le World Wide Web Consortium (W3C). Il s'agit du protocole que l'on utilise au quotidien sans s'en rendre en compte. Dans les langages de programmation, de manière opérationnelle, les métadonnées correspondent à des marqueurs ou balises que l'on introduit dans le code.



Exemple de métadonnées. Source⁴

⁴<https://blog.martignoni.net/2019/09/qu-est-ce-qu-une-metadonnee/>

La structuration des données s'appuie sur des normes, des protocoles et des nomenclatures. Ces référentiels se veulent souvent le plus international possible, pour garantir une interopérabilité optimale entre les systèmes de santé. De nombreuses initiatives ouvertes, comme souvent dans la communauté informatique, ont ainsi vu le jour pour harmoniser les processus de structuration. C'est le cas de schema.org, une communauté dédiée à la promotion des schémas de structuration de pages web. En effet, les données structurées sont également un moyen d'organiser les informations d'une page web, pour en améliorer la clarté et les performances en termes de SEO (Search Engine Optimization), c'est-à-dire la propension d'une page internet à sortir dans les premiers résultats d'une recherche sur un navigateur en fonction de certains mots-clés. La manière dont les résultats de recherche fonctionnent est expliquée en détails par Google dans ses directives et devrait être une lecture de chevet de certains scientifiques pour mieux communiquer et propager leurs résultats.⁵

Prenons l'exemple cité par schema.org d'un site de cuisine qui propose une recette d'un gâteau à la banane.

Sans structuration, le code HTML du site ressemblerait à ceci :

```
1. La fonction gâteau à la banane de ce site  
2. <h1>Gâteau à la banane</h1>  
3. Cette recette consiste en deux étapes : la cuisson et la décoration.  
4. Temps de préparation : 15 minutes  
5. Temps de cuisson : 1 heure  
6. Nombre de personnes : 10 personnes  
7. Type : Gâteau moelleux  
8.  
9. Informations nutritionnelles  
10. 200 (kcalories), 4 grammes de matières grasses  
11.  
12. Ingrédients :  
13.  
14. 3 œufs, 100 grammes de sucre, 100 grammes  
15. 1 œuf  
16. 1 œuf  
17.  
18. 100 grammes de sucre  
19.  
20.  
21. Instructions :  
22. Préchauffer le four à 180 degrés. Mélanger les ingrédients dans un bol. Ajouter la farine en dernier, verser le mélange dans un moule à 9  
23.
```

⁵ <https://developers.google.com/search/docs/fundamentals/seo-starter-guide?hl=fr>

En plus de ces marqueurs, pour considérer une donnée comme structurée, il faut qu'elle soit incluse dans une base de données. Il s'agit d'une collection organisée de données, qui peut aller d'un simple fichier stocké sur un ordinateur personnel jusqu'aux clusters de serveurs organisés en baies de stockage comportant plusieurs couches de sécurité, à la fois physique et informatique. La définition juridique plus large est donnée par l'article L.112-3 al.2 du Code de la Propriété Intellectuelle⁷ : « On entend par base de données un recueil d'œuvres, de données ou d'autres éléments indépendants disposés de manière systématique ou méthodique et individuellement accessible par des moyens électroniques ou par tout autre moyen ».

2. Les modèles de bases de données :

Les systèmes de bases de données occupent une place centrale dans le paysage informatique moderne, notamment dans le domaine de la santé. Deux principaux types de bases de données sont prépondérants : les bases de données relationnelles et non relationnelles. Les bases de données relationnelles, qui reposent sur une variation du modèle proposé par Edgar F. Codd en 1970⁸, utilisent des schémas structurés et préconçus, où les informations sont stockées dans des tables aux relations définies. Les systèmes de gestion de bases de données relationnelles comme MySQL⁹, PostgreSQL ou Oracle, dominent le secteur grâce à leur fiabilité et leur robustesse. Cependant, ils peuvent être moins flexibles face aux changements de structure de données et à la gestion des données semi-structurées ou non structurées et ont du mal à permettre de gérer efficacement de grands nombres de données.

En matière de santé, le choix du type de base de données peut avoir des implications significatives. Dans un contexte où l'exploitation des données de santé est en plein essor, il est probable que le paysage des bases de données continue à évoluer pour répondre à de nouveaux défis. La coexistence et l'intégration de bases de données relationnelles et non relationnelles sont non seulement possibles mais aussi de plus en plus courantes dans les environnements informatiques modernes. Cette approche hybride, souvent désignée par le

⁷ https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006278879

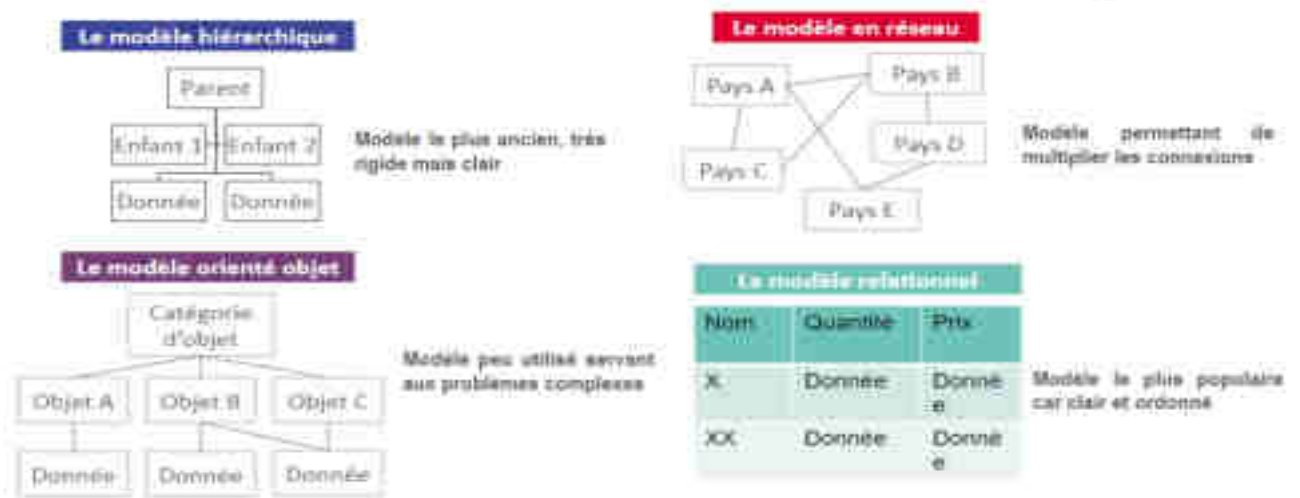
⁸ <https://www.lucidchart.com/pages/fr/schema-base-de-donnees>

⁹ <https://www.mysql.com/fr/>

terme "Polyglot Persistence", consiste à choisir le type de base de données le plus adapté à chaque type de données et à chaque besoin fonctionnel.^{10 11}

Dans un contexte de santé par exemple, une base de données relationnelle peut être utilisée pour stocker des dossiers médicaux structurés, tandis qu'une base de données non relationnelle peut être utilisée pour stocker et traiter des données volumineuses et moins structurées, comme des données génomiques, des images médicales, des notes cliniques en texte libre, etc. L'objectif est d'exploiter les points forts de chaque type de base de données.

Les différents modèles de bases de données



Quelques exemples non-exhaustifs de bases de données

Chacune de ces architectures présente des avantages et des inconvénients en fonction des exigences et des contraintes spécifiques du système de santé concerné. L'adoption d'une architecture adéquate et la mise en place de normes et protocoles communs sont essentielles pour garantir l'interopérabilité, la sécurité et l'efficacité de l'utilisation des données structurées en santé. On peut même aller plus loin en affirmant qu'il est nécessaire de penser à cette intégration dès les premières étapes de la collecte et de la conception de l'architecture des données.

Comme nous le verrons plus loin, c'est l'intégration de données structurées et non structurées qui peut permettre une prise en charge optimale du patient, par exemple en

¹⁰ <https://sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/coronavirus/tout-savoir-sur-le-covid-19/article/tousanticovid>

¹¹ <http://arxiv.org/abs/2204.05779>

combinant les données du dossier médical du patient (données structurées) avec les notes du médecin (données non structurées).

3. La donnée de santé

La donnée de santé, quant à elle, a une définition plus juridique, établie en Europe par le Règlement Général sur la Protection des Données Personnelles (RGPD)¹², mis en application le 25 mai 2018.¹³ Bien que la notion de données de santé soit à apprécier au cas par cas, la Commission Nationale de l'Informatique et des Libertés (CNIL) rappelle que certaines données de santé le sont par nature, notamment des informations telles que les antécédents médicaux, les résultats biologiques et les prescriptions médicamenteuses. D'autre part, certaines données croisées entre elles peuvent ainsi communiquer une information médicale, comme une mesure de poids associée au nombre de pas quotidiens réalisés par une personne. (Art. 2 de la loi Informatique et Libertés dans sa rédaction de 2017 et art. 4 du RGPD).¹⁴

Il existe également des données de santé en raison de la finalité qu'un responsable de traitement peut en faire, c'est-à-dire l'utilisation concrète qui en est fait d'un point de vue médical par un tiers. La donnée de santé fait partie des données à caractère personnel, et toute violation, c'est-à-dire qu'un accès, un effacement ou une modification illégale peut amener des répercussions juridiques. (article 17 du RGPD) Ainsi, il est essentiel pour tous les acteurs de mettre en place une compliance réglementaire adéquate pour assurer la protection, la confidentialité et la sécurité des données.

Les données des patients sont conservées pendant une longue période, généralement 20 ans¹⁵ après la dernière consultation médicale. Cela implique que l'infrastructure qui soutient l'archivage et la sécurité de ces données doit être suffisamment robuste pour conserver l'intégrité, l'anonymisation le cas échéant et la lisibilité des données pendant cette durée. De plus, à tout moment du cycle de vie de cette donnée, un patient peut demander l'application

¹² <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037085952>

¹³ <https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante>

¹⁴ <https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460>

¹⁵ https://www.cnil.fr/sites/cnil/files/atoms/files/referentiel_-_traitements_dans_le_domaine_de_la_sante_hors_recherches.pdf

de certains droits conférés par la loi Informatique et Libertés et renforcés par le RGPD (article 21 du RGPD).

Les objectifs de la structuration des données appliquée au domaine de santé sont multiples. Les outils informatiques qu'utilisent les soignants et chercheurs au quotidien souffrent souvent d'un manque d'ergonomie, de fiabilité et d'interopérabilité. Faciliter l'accès et l'exploitation des informations pour les professionnels de la santé, afin d'améliorer la qualité des soins, la recherche médicale et l'allocation des ressources médicales. Cela peut être réalisé grâce à l'optimisation du SEO par l'utilisation de données structurées : par exemple, les scientifiques qui rédigent des articles pertinents peuvent apprendre à structurer correctement leur contenu pour optimiser leur SEO, en répétant certains mots-clés ou concepts. On peut donc construire une première directive : éduquer les chercheurs et les institutions pour que leurs articles aient un meilleur référencement que des articles pseudo-scientifiques. Une pharmacie peut améliorer sa visibilité sur Internet de cette manière, de même qu'un médecin de ville en liant des articles de santé qu'il aurait écrit sur son site. La mission du professionnel de santé aujourd'hui passe également par l'information dans le web. 7% des recherches Google étant reliées à la santé, c'est-à-dire plus d'un milliard par jour, il est important que ces recherches soient de qualité. ^{16 17}

4. Les niveaux de structuration

La différence entre les données structurées, semi-structurées et non structurées réside principalement dans leur organisation, leur format et leur facilité de traitement par les machines, en particulier dans le contexte de la santé. Les données structurées sont organisées selon une architecture rigide et précise, comme les modèles relationnels, hiérarchiques ou orientés objet. Elles sont généralement stockées dans des bases de données ou des tableaux, avec des colonnes et des lignes clairement définies. Les données structurées en santé incluent des informations telles que les données démographiques des patients, les résultats de laboratoire ou les prescriptions médicamenteuses. Leur structure

¹⁶ <https://www.statista.com/topics/1001/google/#dossier-chapter7>

¹⁷ <https://www.beckershospitalreview.com/healthcare-information-technology/google-receives-more-than-1-billion-health-questions-every-day.html>

préétablie facilite grandement leur traitement, leur analyse et leur intégration dans des systèmes informatiques.

Les données semi-structurées, quant à elles, présentent une organisation intermédiaire entre les données structurées et non structurées. Elles ne suivent pas une architecture aussi rigide que les données structurées, mais contiennent néanmoins des éléments permettant de les organiser et de les interpréter. Les formats XML et JSON illustrent bien ce concept, leur utilisation de balises et de paires clé-valeur dans le code leur conférant une structure à la fois hiérarchique et flexible. Dans le domaine de la santé, les données semi-structurées peuvent inclure des documents annotés avec des métadonnées ou des feuilles de calcul avec des données hétérogènes. Le traitement et l'analyse des données semi-structurées peuvent nécessiter des techniques plus avancées, telles que le traitement du langage naturel (NLP).

Pour finir sur les différentes catégories, les données non structurées sont celles qui ne présentent pas de structure prédéfinie ou de format spécifique, rendant leur traitement et leur analyse par les machines plus complexes. Les données non structurées en santé englobent une grande variété de types de données, comme les notes médicales sous forme de texte libre par exemple. Leur hétérogénéité et leur complexité nécessitent souvent des techniques d'analyse plus élaborées, telles que l'analyse sémantique pour les textes.

En reprenant la définition des niveaux de structure de la donnée que nous avons établie, nous pouvons préciser des exemples concrets et appliqués à la santé :

Données de santé non structurées :

- Des courriels entre professionnels de santé et patients
- Des publications de patients sur les réseaux sociaux ou dans des forums en ligne concernant leur état de santé
- Un compte rendu d'une consultation pris intégralement par écrit

Données de santé semi-structurées :

- Les comptes rendus médicaux d'un logiciel qui contient une partie à choix multiples standardisés et validés par protocole par l'hôpital, tout en incluant une partie en texte libre pour que le soignant puisse ajouter ses précisions.

- Des résultats d'imagerie médicale archivés et stockés, contenant des annotations de la part du soignant

Données de santé structurées :

- Des résultats de tests de laboratoire indexés sur le LOINC
- Actes médicaux codés en nomenclature type CCAM.

Données hyper-structurées :

- Des résultats de tests de laboratoire indexés sur le LOINC, requêttables via une interface utilisateur ergonomique, stockés dans la base de données sécurisée interne, interopérable et interfacée avec les hôpitaux.
- Données d'admission et de sortie d'hôpital intégrés dans un DPI capable de gérer les séjours en accord avec les règles établies par la T2A. Ces données intègrent des actes médicaux codés en nomenclature type CCAM.

Chez Galeon, sans forcément en réclamer la paternité, nous parlons de données hyperstructurées. Si le premier défi est véritablement de ranger les données de manière optimisée pour le soignant et le patient, ainsi que de faire de la recherche en requêtant facilement parmi ces dites données, on peut aller plus loin. Connecter des institutions généralement peu interopérables entre elles, avec des données robustes, de haute qualité et facilement requêttables est notre objectif principal.

Ces distinctions sont d'autant plus importantes que tous ces exemples produisent de la donnée de santé importante au quotidien pour la prise en charge du patient, et peuvent guider la conduite à tenir du personnel médical. Pour autant, toute donnée ne gagne pas forcément à être structurée. Permettre aux patients de s'exprimer sur leur maladie, témoigner, relève de la liberté individuelle et peut tout à fait aider aux soins. Une indication non structurée en texte libre glissée par un soignant dans un dossier médical peut devenir salvatrice si elle communique une information absente du logiciel ou du protocole, car chaque situation clinique est différente.

En ce sens, une revue de la littérature amène à la conclusion que c'est bien la combinaison des données structurées et non structurées mises en relation par l'équipe soignante qui

forme l'information la plus complète au sens que nous avons défini plus haut.^{18 19} Cet article explore la possibilité de combiner des données structurées et non structurées dans les dossiers patients informatisés (DPI) pour créer des cohortes de patients possédant certains critères d'inclusion. Les auteurs ont identifié les patients atteints de syndrome coronarien aigu dans deux districts de santé en Australie. Ils ont combiné des critères d'inclusion basés sur des données structurées (codes SNOMED CT et CIM-10 pour l'infarctus aigu du myocarde) et des données non structurées (recherche de mots-clés dans les champs de texte libre).

L'étude a révélé que l'utilisation de données non structurées en combinaison avec des données structurées a permis d'identifier un plus grand nombre de patients atteints de syndrome coronarien aigu que l'utilisation de données structurées seules. Plus précisément, l'introduction de données non structurées a permis d'identifier 5,4 fois plus de patients atteints de syndrome coronarien aigu que l'utilisation de codes SNOMED CT et CIM-10 seuls. De plus, l'ajout de données non structurées a permis d'identifier des patients qui étaient plus susceptibles d'être hospitalisés pour un syndrome coronarien aigu, par des précisions cliniques qui n'étaient pas prévues par l'éditeur du logiciel de manière structurée. Le logiciel utilisé pour cette étude n'était pas à la hauteur en termes de structuration et d'optimisation de la donnée, d'où l'importance de développer des outils puissants en ce sens.

Citons également cet article, dans lequel les auteurs présentent et explorent des méthodes pour utiliser à la fois les éléments de données structurées et les descriptions en texte libre dans des rapports d'incidents sanitaires pour identifier et classer les événements similaires.

²⁰ Les différentes techniques de recherche et de pondération améliorent potentiellement la compréhension des agents de sécurité et d'autres professionnels de la santé sur les rapports d'incident liés.

¹⁸ <https://doi.org/10.1038/s41746-023-00817-8>

¹⁹ <https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01441-w>

²⁰ <https://www.sciencedirect.com/science/article/pii/S153204641500204X>

5. *Revue des directives pour structurer, acquérir, extraire et normaliser la donnée*
Au-delà de la classification en fonction du degré de structure de la donnée, on peut également parler du paramètre de qualité de la donnée. Dans un cadre scientifique, le fait d'avoir des données partageant des caractéristiques spécifiques facilite leur réutilisation, permet d'éviter certains biais et simplifie la révision par les pairs des hypothèses avancées dans les publications. Il y a un véritable problème de valorisation des données. C'est pour répondre à ce besoin que les principes FAIR ont été introduits et publiés dans la revue *Scientific Data* en 2016 ²¹. Le 'F' dans FAIR signifie 'Findable', impliquant que les données doivent être facilement repérables et localisables par les êtres humains et les machines. Cela peut être réalisé en attribuant des identifiants uniques et persistants aux données, en utilisant des métadonnées pour décrire les données et en rendant les données accessibles via les moteurs de recherche et les catalogues.

De plus, il faut veiller à l'accessibilité de ces données. Les données doivent être ouvertement accessibles à tous, sans aucune restriction ni barrière. Cela peut être réalisé en fournissant un accès ouvert aux données, en utilisant des formats standard et interopérables et en fournissant une documentation claire et concise. Un élément de plus en plus important est évidemment l'interopérabilité des données. Elles doivent pouvoir être combinées et intégrées avec d'autres données provenant de différentes sources. Cela peut être réalisé en utilisant des formats standard et interopérables, en adoptant des vocabulaires et des ontologies communs et en fournissant une documentation claire et concise. Enfin, comme dit plus haut, ces données doivent être réutilisables. En effet, elles doivent pouvoir être réutilisées à des fins différentes et par différents utilisateurs. Cela peut être réalisé en fournissant une documentation claire et concise, en utilisant des formats standard et interopérables et en fournissant un accès ouvert aux données.

En France, ces principes ont été largement acceptés et diffusés. Ainsi l'ANR l'utilise dans des projets et adopte une stratégie de science ouverte. ^{22 23}

²¹ <https://www.nature.com/articles/sdata201618>

²² <https://anr.fr/fr/lanr/engagements/la-science-ouverte/?mdrv=anr.fr>

²³ <https://openscience.pasteur.fr/2020/10/05/les-principes-fair-findable-accessible-interoperable-reusable>

La manière de rentrer les données sur un logiciel, lui-même répondant au maximum aux normes de structuration, est également très importante. En effet, les outils restent opérateurs dépendants et il n'est pas question d'agents autonomes récoltant les données de soins seuls, encore moins avec l'IA Act à l'horizon.²⁴ Lors de la collecte de données, il est important de garantir la qualité et la précision des informations enregistrées. Cela peut être réalisé en utilisant des protocoles de saisie des données rigoureux, en formant le personnel sur les meilleures pratiques de documentation et en mettant en place des mécanismes de contrôle de la qualité des données. Il est également primordial, à la base, que l'outil informatique utilisé soit bien conçu et ergonomique. L'interface doit être suffisamment attrayante et ne pas entraîner de la fatigue de l'utilisateur du fait de trop nombreux clics pour atteindre la fonction désirée. Les champs structurés doivent interagir entre eux, se préremplir ailleurs s'ils ont bien été remplis une première fois, et servir à des requêtes pour faciliter la recherche sur les données rentrées.

Pour extraire des données structurées à partir de sources non structurées ou semi-structurées, il existe des techniques telles que le traitement automatique du langage naturel (NLP) et les algorithmes de machine learning. Pour les données textuelles, il est possible d'extraire des mots sous forme d'entités, des relations et des concepts clés et d'entraîner des algorithmes de machine learning pour classifier, extraire et structurer plus facilement des données.

On peut citer quelques techniques courantes de NLP, comme la tokenisation qui consiste à diviser un texte en unités plus petites, appelées tokens, qui peuvent être des mots, des phrases ou des phrases plus longues. ChatGPT, par exemple, découpe nos requêtes en tokens que l'algorithme traite de manière probabiliste.²⁵ On pourra ensuite procéder à une analyse morphologique, qui permet d'analyser la structure des mots en identifiant leur racine (ou radical) et leurs affixes (préfixes, suffixes, infixes). L'analyse morphologique est utile pour la reconnaissance des mots et la réduction de la taille du texte par exemple. Un domaine très intéressant et en pleine ébullition est l'analyse du sentiment, qui vise à déterminer l'opinion, l'émotion ou l'attitude exprimée dans un texte. Elle peut être utilisée pour analyser des commentaires, des critiques ou des publications sur les réseaux sociaux

²⁴ <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52021PC0206>

²⁵ <https://www.ibm.com/topics/natural-language-processing>

afin de comprendre les opinions des utilisateurs sur un produit, un service ou un événement. Cela peut également servir à l'éthique ou la censure pour un algorithme, qui peut refuser de répondre à une requête dangereuse ou formulée de manière très impolie.

Les données, encore plus celles médicales et sensibles, nécessitent d'être normalisées, standardisées. Cela peut inclure la conversion des unités de mesure, la gestion des données manquantes et la résolution des ambiguïtés. Un exemple simple est l'utilisation du point comme séparateur décimal dans les pays anglo-saxons alors que nous utilisons la virgule en France. Cela peut sembler peu mais cette nuance peut mener à des erreurs d'interprétation dans certains résultats biologiques et ne pas être pris en compte dans l'harmonisation des pratiques. La normalisation facilite la comparabilité, l'analyse et l'échange des données de santé.

La gestion des données de santé structurées implique la mise en place de procédures de maintenance, de sauvegarde et de récupération des données pour assurer la disponibilité et la durabilité des informations médicales. Il est également essentiel de surveiller et d'évaluer régulièrement la qualité et l'intégrité des données pour détecter et résoudre les problèmes éventuels. La mise en place d'une infrastructure technique pour les données structurées de santé doit inclure la formation et le soutien des professionnels de la santé, des informaticiens et des administrateurs pour assurer une utilisation appropriée et efficace des systèmes.

Outre la structuration des données, il existe également tout un champ dédié à l'annotation et l'indexation des données. Cela peut inclure l'assignation de balises, de mots-clés ou de métadonnées structurées aux données non structurées afin de les rendre plus facilement interrogeables. Des entreprises et des algorithmes ont été créés autour de l'annotation de données qui est un domaine qui travaille en synergie avec la structuration. La différence réside dans le fait que la structuration est une caractéristique intrinsèque de la donnée alors que l'annotation est un processus postérieur à l'entrée de la donnée.



L'annotation de données. Source²⁶

3. Exemples de nomenclatures

La Classification Internationale des Maladies (CIM-11) et la SNOMED CT sont deux systèmes de codification et de classification largement utilisés dans le domaine de la santé pour structurer les données médicales.

a. Classification Internationale des Maladies (CIM-11 ou ICD-11) :

La CIM-11 est une classification développée par l'Organisation mondiale de la santé (OMS) pour classer les maladies, les troubles, les symptômes, les causes de décès et les facteurs influençant l'état de santé. Arrivée en janvier 2022, sa dernière version est sortie en janvier 2023 et vient remplacer une CIM-10 vieillissante adoptée en 1990.²⁷ Notons qu'en France, on utilise actuellement la CIM-10 modifiée du fait de sa forte propagation dans les centres de soins et l'affinage qui s'est fait autour au fur et à mesure des années. J'ai choisi de parler de la CIM-11 pour tourner mon propos vers le futur.²⁸ La CIM-11 décompose chaque diagnostic en une série de codes alphanumériques qui correspondent à des catégories et sous-catégories de diagnostics distincts qui prend en compte les synonymes. Par exemple, en tapant mucoviscidose, il sera proposé le codage CA25.0 Fibrose kystique, qui est le nom CIM-11. Chaque maladie est ainsi clairement identifiée et la CIM-11 est utilisée à l'hôpital et

²⁶ <https://fr.shaip.com/blog/the-a-to-z-of-data-annotation/>

²⁷ <https://www.cepidc.inserm.fr/causes-medicales-de-deces/classification-internationale-des-maladies-cim?undefined>

²⁸ <https://www.who.int/news/item/14-02-2023-icd-11-2023-release-is-here>

en ville pour aider au codage et à la facturation des actes médicaux, en permettant l'attribution d'un diagnostic principal et secondaire pour chaque séjour hospitalier par exemple. Elle est utilisée en complément des classifications CCAM, NGAP pour les consultations médicales et LOINC pour les actes biologiques. En plus de cela, la CIM-11 permet de codifier les antécédents médicaux des patients dans une norme qui se veut universelle selon l'OMS et qui peut donc être comprise de tous les soignants, même à l'international. Ces nomenclatures partagées contribuent à l'harmonisation des pratiques.²⁹. Le guide d'implémentation de la CIM-11 précise cependant qu'en cas de nécessité d'un terme spécifique à leur pratique, « les pays pourront ajouter des éléments au Guide de référence, en fonction des normes ou lignes directrices devant être appliquées pour la notification au niveau national, tout en veillant à la cohérence internationale du codage et de la présentation des statistiques ».

La CIM-11 et sa version antérieure CIM-10 sont utilisés pour 70 % des dépenses mondiales de santé aux fins du remboursement et de l'allocation des ressources et concernent 60 % de la population mondiale dans plus de 110 pays selon les chiffres de l'OMS. On constate donc une implémentation très forte de cette norme.

La CIM-11 est à la fois une nomenclature et un ensemble d'outils informatiques pour aider les soignants sans rajouter une surcouche administrative lourde. L'OMS propose de nombreux outils pour intégrer l'API de la CIM-11 dans un système informatique et peut être utilisée hors-ligne et en ligne. La CIM-11 permet d'obtenir de la donnée standardisée pour établir des statistiques de morbidité et de mortalité au sein du réseau d'établissement qui l'opère. Elle démontre un intérêt en épidémiologie et en pharmacovigilance. De plus, l'établissement de bases de données patients en CIM-11 présente un intérêt en recherche. On peut tout à fait imaginer étudier avec beaucoup de données, les fréquences d'apparition de plusieurs CIM-11 et donc découvrir des liens entre des troubles que l'on avait jusqu'alors pas soupçonnés, faute de données de qualité ou quantité suffisantes.

²⁹ 11_Implementation_or_Transition_Guide_rev_25_10_2019-fr.pdf



La France utilise la CIM-10 depuis 2000. L'OMS propose dans son guide une période de transition de 2-3 ans vers la CIM-11 pour les pays intégrant la CIM-10, il est donc important que la France prenne le virage à temps pour produire des données à jour. En ce sens, la traduction complète de la CIM-11 a été assurée par l'Agence Nationale de la Santé (ANS).

La CIM-11 a été conçue pour venir compléter les logiciels et les systèmes informatiques existants à l'appui du dossier médical électronique. Les éléments d'interopérabilité intégrés à la CIM-11 améliorent les possibilités de récupération des données et l'intégration d'autres fonctions, telles que les données pharmaceutiques et de laboratoire. Les éditeurs de logiciels de santé doivent donc se mettre à jour pour intégrer cette nouvelle classification.

On comprend donc la nécessité de l'acculturation à l'utilisation de ces nomenclatures en France dans le milieu médical, pour favoriser l'harmonisation des terminologies, des pratiques et l'interopérabilité.

b. SNOMED CT (Systematized Nomenclature of Medicine - Clinical Terms) :

SNOMED CT est une terminologie médicale développée par SNOMED International, une organisation à but non lucratif. Cette organisation, qui peut être rejointe par des pays en tant que pays membres ou affiliés, a une vision claire. Ainsi selon leur présentation « A case

for investment », en 2025 les terminologies cliniques ont pour ambition d'être utilisées mondialement au travers d'un langage de santé unique.³⁰ On remarque à nouveau une vision universelle de la nomenclature de la santé.³¹

C'est une terminologie clinique qui se veut la plus exhaustive possible et multilingue. La SNOMED-CT consiste en une base de données de plus de 350 000 concepts cliniques qui couvrent un large éventail de domaines médicaux, tels que des diagnostics, des procédures, des dispositifs médicaux et des substances. Tout cet ensemble est interconnecté et produit plus de 3 200 000 relations entre ces items. Ces relations peuvent par ailleurs être passées au peigne fin, comme dans cette étude américaine publiée dans le Journal of Bioinformatics en mars 2023³². Les auteurs ont défini des distances entre des paires de maladies à l'aide de la CIM-10, en la croisant avec la SNOMED CT. La distance est représentée graphiquement et est comparée à d'autres métriques comparatives entre les maladies, mettant en évidence des liens plus forts et insoupçonnés entre certaines maladies qui pourront potentiellement servir de base de futures recherches. La conclusion montre que la construction de représentations spatiales à partir des données de bases de classification comme la CIM-10 ou la SNOMED CT peut ouvrir des champs thérapeutiques.

Utiliser la SNOMED CT facilite l'enregistrement, la communication et l'échange d'informations médicales structurées entre les professionnels de la santé et les institutions, ce qui en fait un outil central dans les bases structurées des données de santé.

La SNOMED-CT se veut être une classification ontologique, c'est-à-dire qu'elle est une source pour d'autres classifications et définit les concepts médicaux en tant que tels, de manière presque philosophique. Toute une galaxie de terminologies gravite autour de la SNOMED-CT, comme le montre cette image.

³⁰ <https://esante.gouv.fr/espace-presse/interoperabilite-semantique-la-france-choisit-la-snomed-ct-pour-la-description-des-localisations-anatomiques>

³¹ https://icd.who.int/fr/docs/192190_ICD-11_Implementation_or_Transition_Guide_rev_25_10_2019-fr.pdf

³² <https://doi.org/10.1016/j.jbi.2023.104307>



Du côté de la France, un accord d'accès anticipé a été signé le 10 novembre 2022 entre SNOMED International et l'ANS (Agence Nationale en Santé), précédemment ASIP Santé, en attendant l'adhésion finale en seconde étape qui sera co-financée par l'Union européenne.

Cette décision est en accord avec la construction de l'espace européen de données de santé (EHDS), axe stratégique français au niveau de la donnée de santé.³³

Ce contrat soulève des questions de sécuritaires et juridiques, car il faut que les rapports annuels envoyés à la SNOMED International respectent le RGPD et que les données soient sécurisées en cas de résiliation.

À ce jour il n'existe pas encore de version officielle en français de la SNOMED-CT, bien qu'une version belge, déjà disponible, puisse aider les soignants francophones.³⁴ Cet accord permettra à la SNOMED-CT de faire partie de la discussion de la vague 2 du Ségur de la Santé qui se tiendra en septembre 2023, et incitera à équiper les solutions informatiques qui se certifieront.

4. Le DPI

Dans le paysage actuel des systèmes de santé, le Dossier Patient Informatisé (DPI), ou Electronic Health Record (EHR) en langue anglaise, se présente comme un dispositif numérisé de centralisation et de gestion des informations médicales propres à chaque

³³ https://ec.europa.eu/commission/presscorner/detail/fr/ip_22_2711

³⁴ <https://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-snomed-ct>

patient. On utilisera ce terme de manière interchangeable avec Dossier Médical Informatisé, ou DMI, la nuance entre les deux étant légère. Les DMI sont décrits de manière légale dans la loi du 4 mars 2002 relative aux droits des malades.³⁵ Cet outil trouve une place de choix au sein des Systèmes d'Information Hospitaliers (SIH) contemporains, en quête d'une optimisation de l'accessibilité aux données médicales. De très nombreux DPI sont en concurrence en France, ce qui participe à l'innovation mais en contrepartie rend parfois difficile l'interopérabilité des SIH. Changer de DPI est un processus long et nécessite une équipe projet dédiée, ainsi qu'un budget. L'idée principale est de faciliter l'élaboration d'un plan de soins sur-mesure et efficient, grâce à la collecte précise et exhaustive de données. Tout l'intérêt se situe dans le gain de temps réalisé par le soignant à ne pas inutilement remplir des informations qui auraient déjà pu être collectées par le passé pour un patient donné, ou présenter les informations clés de manière claire pour éviter de les chercher trop longtemps. Cela peut parfois paraître anecdotique mais des gains de quelques minutes par patient peuvent s'additionner en des heures de soins réels auprès d'un patient, ce qui est substantiel. Il n'y a rien de plus frustrant pour un soignant en fin de garde de devoir remplir en urgence, à la fin de son service, des informations patient parce que son outil n'est pas ergonomique.³⁶

Le DPI donne accès à une vision globale de l'historique du patient, couvrant les antécédents médicaux et chirurgicaux, les prescriptions de médicaments, les allergies identifiées, sans oublier les résultats d'examens de laboratoire et d'imagerie. Ces données, continuellement mises à jour, sont aisément accessibles aux professionnels de santé autorisés, ce qui fluidifie le processus de diagnostic et la prise de décisions thérapeutiques. La notion d'interopérabilité est donc primordiale, permettant de contourner les obstacles rencontrés lorsqu'un patient dispose de données cliniques, comme une radiographie, stockées dans un système spécifique à un établissement de santé et qui ne peut être envoyée à temps à un autre hôpital concourant aux soins du patient. Comme souligné dans une étude américaine, le manque de données structurées et fiables peut freiner la mise en œuvre de projets

³⁵ <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000227015/>

³⁶ <https://www.nouvelobs.com/rue89/rue89-temoignage/20130523.RUE6386/a-montpellier-les-degats-de-l-hopital-numerique.html>.

hospitaliers, renforçant l'importance de construire des bases de données durables et facilement navigables.³⁷

L'avènement de l'Intelligence Artificielle (IA) et du Machine Learning a marqué une étape supplémentaire dans l'évolution des DPI, en offrant la possibilité de détecter des tendances et des modèles à même les données des patients. Ces avancées contribuent à une amélioration du suivi des patients et à une analyse plus fine des tendances démographiques, à condition toutefois de relever des défis non négligeables en matière de confidentialité et de sécurité des données. La protection des données sensibles des patients et le RGPD est d'ailleurs au cœur des préoccupations des régulateurs, qui sont appelés à instaurer des normes de sécurité rigoureuses. Par ailleurs, la standardisation des données entre les systèmes de DPI variés, et l'interopérabilité entre les différents systèmes de santé sont des questions cruciales à résoudre pour maximiser l'efficacité du DPI.

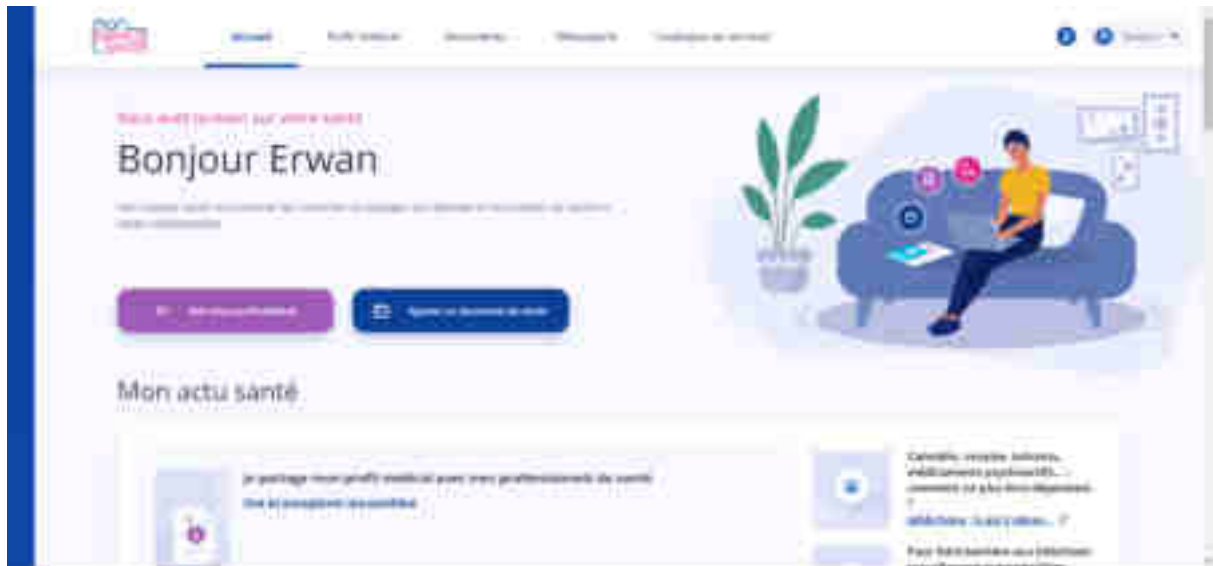
Le choix a été laissé en France aux acteurs privés la possibilité de développer leurs propres systèmes de DPI et technologies, tout en conditionnant l'intégration de ces données au sein du Dossier Médical Partagé (DMP). Face à ces défis, l'harmonisation et la structuration des normes apparaissent comme des enjeux majeurs pour l'exploitation efficace d'un grand volume de données. Un DPI national partagé avec les patients pourrait éventuellement voir le jour, comme en Suède par exemple.³⁸

5. Mon Espace Santé et le DMP

Mon Espace Santé est une plateforme informatique et l'un des volets clés du Ségur du Numérique. Elle a été conçue pour optimiser la gestion des informations de santé des individus en favorisant une approche centrée sur le patient au niveau national. Cette solution numérique se manifeste par un portail en ligne sécurisé permettant aux citoyens de centraliser, gérer et consulter leurs informations de santé de manière autonome et éclairée.

³⁷ <https://doi.org/10.1287/inte.2020.1036>

³⁸ <https://www.cairn.info/revue-francaise-des-affaires-sociales-2020-1-page-313.htm?ref=doi>



Le prédécesseur de Mon Espace Santé était le Dossier Médical Personnel (DMP), qui a été lancé en 2011. Le DMP était un service en ligne qui permettait aux patients de stocker et de partager leurs informations de santé avec les professionnels de santé. Cependant, le DMP n'a pas connu le succès escompté et a été remplacé par Mon Espace Santé en 2019.

De nouvelles critiques ont rapidement émergé sur cette mouture. Tout d'abord, il était perçu comme un empilement de documents disparates et difficile à consulter, ce qui rendait son accès complexe pour les patients et les professionnels de santé. De plus, des problèmes techniques et des retards dans son déploiement ont suscité des frustrations et un manque de confiance chez les utilisateurs. Certains professionnels de santé ont exprimé des réticences à utiliser le DMP nouvelle version en raison de préoccupations liées à la charge de travail supplémentaire, à la confidentialité des données et à la compatibilité avec leurs systèmes existants. De plus, une communication insuffisante a limité la connaissance et l'adoption du système par les patients et les professionnels de santé. Enfin, le coût élevé de mise en place et de maintenance de Mon Espace Santé a dissuadé certaines parties prenantes de s'engager pleinement dans son utilisation selon un rapport officiel de la Cour des Comptes.³⁹ Ces facteurs ont contribué à l'échec relatif du DMP, qui fut remplacé par Mon Espace Santé., lancé pour remédier à ces problèmes et a connu une croissance rapide depuis son introduction.

³⁹ https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/134000136.pdf

En intégrant l'ensemble des données de santé d'un individu, Mon Espace Santé offre un accès simplifié à divers types d'informations, telles que les résultats d'analyses médicales, les prescriptions, les comptes rendus d'hospitalisation et les données de vaccination. La plateforme offre un suivi continu et exhaustif de la santé de l'individu, favorisant ainsi la compréhension et la gestion de sa propre santé.

Mon Espace Santé illustre également l'importance de l'interopérabilité des données de santé. La plateforme permet la consolidation des données provenant de diverses sources, offrant, en théorie, une vision unifiée du parcours de soins d'un individu. Cette intégration se veut d'aller plus loin depuis le projet d'intégration récente du Dossier Pharmaceutique selon un décret récent du 3 avril 2023.⁴⁰ Selon Xavier Vitry, Directeur de projet en charge du volet officine du Ségur : « Pour être effective, la connexion entre le DMP et le DP demande encore des évolutions, sur lesquelles travaille l'Ordre des pharmaciens. Il faudra notamment que le DP soit en conformité avec le format européen de dispensation, et en mesure d'utiliser l'Identité Nationale de Santé (INS). Tout cela va s'aligner progressivement. » Le Dossier Pharmaceutique existe depuis 2007 et est une interface présente dans toutes les officines du territoire national. Il recense les médicaments qui ont été délivrés au cours des 4 derniers mois, ainsi que les traitements et prises en cours. Les médicaments figurant sur le dossier peuvent avoir été prescrits par un médecin ou avoir été achetés en OTC. Les interactions médicamenteuses possibles sont marquées clairement.

Les objectifs du DP, affichés par le gouvernement sur le site du service public, sont :

- Sécuriser la dispensation des médicaments (médecin, pharmacien, hôpital) en limitant les risques d'interactions médicamenteuses et les traitements redondants.
- Garantir une coordination entre les professionnels de santé
- Favoriser l'amélioration de la couverture vaccinale.⁴¹

Du côté des citoyens, un bilan d'Assos Santé indique que « 82 % des Français ont déjà entendu parler de Mon Espace Santé (MES), avec un « bon taux d'activation pour un service

⁴⁰ <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000047396628>

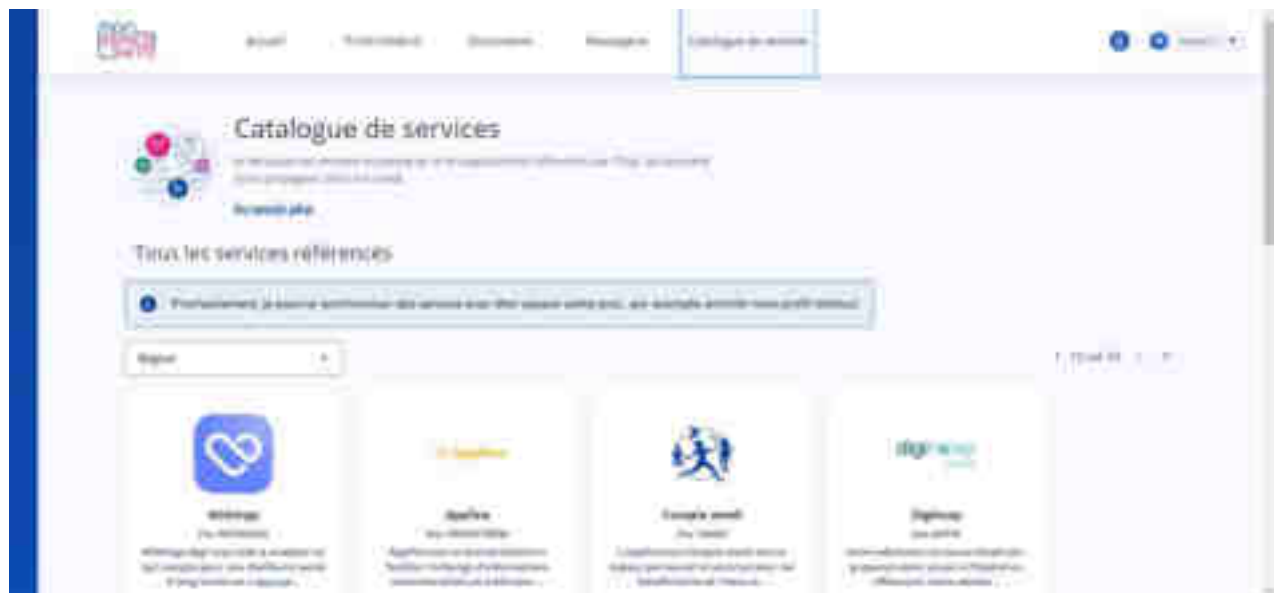
⁴¹ <https://www.service-public.fr/particuliers/vosdroits/F16033>

récent » (38 %), mais des freins restent à lever pour améliorer l'appropriation du service : 45 % des Français n'ont rien fait concernant leur espace santé ». ⁴²

Toujours selon cet article, Mon Espace Santé est un espace jugé simple et ergonomique mais plutôt long à renseigner si l'on veut avoir un dossier complet.

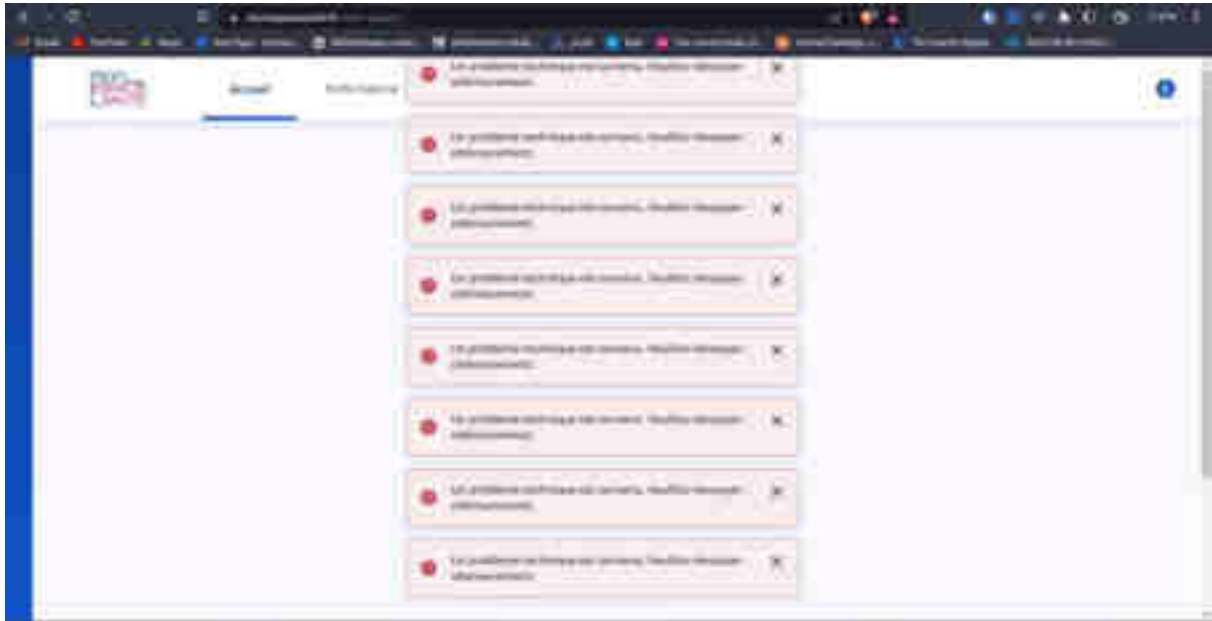
De plus, 5 % des Français auraient fait opposition à la création de leur compte ou supprimé leur compte. Deux tiers de ces citoyens l'auraient fait par manque de confiance concernant les données personnelles.

Mon Espace Santé présente un catalogue d'applications externes qui peuvent s'interfacer et qui permettent de créer un véritable écosystème de santé.



De mon côté, j'ai pu effectivement constater quelques problèmes techniques pour une utilisation limitée :

⁴² <https://www.healthandtech.eu/fr/article/view/15627/espace-sante-bonne-notoriete-mais-freins-appropriation-bilan-france-assos.html>



Mon Espace Santé fait néanmoins des émules à l'international, avec notamment la NHS Scotland qui va lancer une plateforme similaire.⁴³

7. Règlementation en vigueur en France (HDS, RGPD)

En France, la protection des données et la conformité réglementaire sont des enjeux cruciaux dans le domaine de la santé, notamment en raison de la sensibilité des informations médicales et de la nécessité de garantir la confidentialité et la sécurité des données des patients. La Commission Nationale de l'Informatique et des Libertés (CNIL) est l'autorité française chargée de la protection des données personnelles, et elle veille au respect des réglementations en vigueur, comme le Règlement Général sur la Protection des Données (RGPD) au niveau européen. Ce fameux RGPD, entré en vigueur en mai 2018, établit des normes strictes pour la collecte, le traitement, le stockage et la transmission des données personnelles, y compris les données de santé. Il impose des obligations aux responsables de traitement et aux sous-traitants en matière de protection des données, telles que la mise en place de mesures de sécurité appropriées, la notification des violations de données aux autorités compétentes et la désignation d'un délégué à la protection des données (DPO) dans certaines organisations.

⁴³ <https://www.healthandtech.eu/fr/article/view/15578/uk-nhs-scotland-lance-national-digital-platform-equivalent-ecossais-espace.html>

En plus du RGPD, la France dispose de sa propre législation en matière de protection des données de santé, notamment la loi Informatique et Libertés. Cette loi établit des règles spécifiques pour le traitement des données de santé, comme l'obligation de recueillir le consentement explicite des patients pour le traitement de leurs données médicales et la limitation du partage de ces données aux seuls professionnels de santé directement impliqués dans la prise en charge du patient. Une veille juridique approfondie doit être effectuée par les responsables de traitement voulant gérer de la donnée médicale, et la structurer permet d'en connaître plus aisément certaines caractéristiques qui rendent sa qualification juridique plus facile.

Les établissements de santé et les professionnels de la santé en France sont tenus de respecter les normes de sécurité des systèmes d'information, telles que définies par l'Agence du Numérique en Santé (ANS).⁴⁴ Ces normes incluent des exigences en matière de chiffrement des données, d'authentification des utilisateurs, de contrôle d'accès et de sauvegarde des données. Des recommandations sont régulièrement publiées du fait des nombreuses cyberattaques sur les établissements de santé du territoire, la voie d'entrée de ces attaques provenant souvent d'une maladresse interne de la part d'un soignant.

La certification Hébergeur de Données de Santé (HDS) est une exigence spécifique en France pour les fournisseurs qui hébergent des données de santé à caractère personnel. Cette certification a été mise en place pour garantir la protection, la confidentialité et la sécurité des données de santé hébergées. Les hébergeurs de données de santé doivent obtenir cette certification pour être autorisés à stocker et traiter des données médicales en France, et donc les éditeurs de logiciels et hôpitaux utilisant de la donnée structurée ou non.

Les données générées par la recherche impliquant la personne humaine doivent être stockées dans des infrastructures apposées HDS, ou Hébergeur de Données de Santé. Certaines entreprises font confiance à des acteurs externes qui disposent de la certification et stockent leurs données sur les serveurs de ces entreprises, mais la certification HDS peut également être obtenue en interne via une procédure d'audit qu'un organisme accrédité par le COFRAC mandate.⁴⁵ L'audit a deux composantes, une documentaire, où l'auditeur vérifie que

⁴⁴ <https://esante.gouv.fr/produits-services/hds>

⁴⁵ <https://www.cofrac.fr>

l'entreprise a bien rempli tous les documents relatifs à la certification et un sur site, qui dure plusieurs jours. L'audit sur site donne lieu à la délivrance d'un certificat de conformité d'une durée de trois ans, avec un audit de surveillance annuel du fait de la sensibilité des données en cause.

La certification HDS concerne 6 domaines de compétences répartis en deux périmètres. Un certificat « hébergeur d'infrastructure physique » pour les activités de mise à disposition de locaux d'hébergement physique et d'infrastructure matérielle. Il concerne les compétences 1 et 2. Un certificat « hébergeur infogéreur » pour les activités de mise à disposition d'infrastructure virtuelle, de mise à disposition de plateforme logicielle, d'administration/exploitation et de sauvegarde externalisée pour les compétences 3,4,5 et 6. Un hébergeur peut se faire certifier dans un seul de ces certificats ou les deux. Cette distinction est très importante car une entreprise qui souhaite faire appel à une gestion de ses données chez un prestataire externe ayant obtenu l'HDS doit bien vérifier que le prestataire est certifié sur celles des 6 compétences qui concerne les données à protéger.

La certification HDS est une composante primordiale de la stratégie juridique du développement d'un produit pharmaceutique, qui génère de la donnée médicale plus sensible que d'autres catégories. Cette donnée peut être pensée pour être structurée en amont et ainsi faciliter son traitement.

Tous les quatre ans, les établissements de santé publics et privés sont audités par des experts-visiteurs de la HAS.⁴⁶ Cette dernière octroie une certification qui évalue la qualité et la sécurité des soins, ce qui permet à l'établissement d'obtenir l'un des niveaux suivants :



⁴⁶ https://www.has-sante.fr/jcms/c_411173/fr/comprendre-la-certification-pour-la-qualite-des-soins

Obtenir ce badge est un grand enjeu pour les hôpitaux, qui peuvent risquer jusqu'à la fermeture. Posséder des outils permettant de correctement stocker et utiliser des données structurées ne peut qu'aider, notamment dans la récolte et la définition de KPI en temps réel pour la certification (voir 2. Application au domaine de la santé).

8. Entrepôts de santé

En France, il est possible de stocker des données en grand nombre et provenant de sources hétérogènes dans des infrastructures appelées entrepôts. Ces systèmes doivent garantir la sécurité, la confidentialité, l'intégrité et la disponibilité des données, tout en permettant des opérations de requête et d'analyse efficaces. Cela est rendu possible en harmonisant les jeux de données sous un format commun et exploitable. Toute constitution d'un entrepôt de données de santé doit faire l'objet d'une demande préalable à la Commission nationale de l'informatique et des libertés (CNIL), qui a récemment publié un référentiel à ce sujet. Il s'agit en effet d'un traitement de données à caractère sensible dans l'intérêt public, ce qui tombe sous l'article 6.1 du RGPD. Le projet est ensuite analysé par un comité scientifique et éthique.

La Haute Autorité de Santé (HAS) a publié un rapport en 2022 pour établir un panorama des entrepôts de santé en France :

On recense ainsi 22 entrepôts de santé, dont 17 au sein d'un centre hospitalo-universitaire et 5 dans d'autres types d'établissements de santé. Les entrepôts sont principalement situés dans des CHU en raison des coûts élevés associés à leur mise en place. L'effectif qui est affecté à la gestion de ces entrepôts est extrêmement varié et ne fait pas l'objet d'une standardisation à l'échelle nationale. Les entrepôts de santé possèdent un socle commun de données comprenant les données administratives, les textes cliniques, la biologie, le circuit du médicament... et un socle variable de données dans leur format ou leur construction. Les données gérées ainsi que leurs quantités sont donc extrêmement variables.

Un appel à projets de 50M€ a été lancé par le ministère de la santé et de la prévention et en particulier la DGOS.⁴⁷ Cet appel à projets visait à mettre en place et consolider un réseau des EDSH afin de stimuler l'écosystème public et privé de la recherche et de l'innovation en

⁴⁷ <https://www.health-data-hub.fr/aap-eds>

santé et placer la France en leader dans la gouvernance des données de santé. La HAS préconise également dans son rapport quelques axes d'amélioration, comme un élargissement de ces entrepôts à des données non purement hospitalières, comme de la facturation de ville par exemple. Combiner la ville et l'hôpital serait un moyen de construire une gouvernance unifiée de ces données.

La HAS corrobore un axe dont nous avons déjà parlé, qui est que « La mise en place d'une infrastructure technique pour les données structurées de santé doit inclure la formation et le soutien des professionnels de la santé, des informaticiens et des administrateurs pour assurer une utilisation appropriée et efficace des systèmes. »⁴⁸

9. Réglementation aux États-Unis

Aux États-Unis, la protection des données de santé et la conformité réglementaire sont régies principalement par la loi Health Insurance Portability and Accountability Act (HIPAA) de 1996. Cette loi établit des normes nationales pour la protection des informations médicales sensibles et la confidentialité des patients, à l'instar du RGPD qui n'est pas applicable outre Atlantique. Les deux réglementations ont des approches légèrement différentes en matière de protection des données et la portée du RGPD est bien plus grande. Il couvre une variété beaucoup plus large de données personnelles et pas seulement les informations de santé. Elle s'applique à toutes les entreprises qui traitent les données des résidents de l'UE, qu'elles soient basées dans l'UE ou non. De plus, le RGPD donne aux individus un certain nombre de droits supplémentaires par rapport à leurs données, comme le droit à l'oubli, qui n'est pas explicitement présent dans la HIPAA.^{49 50}

Au lieu de données sensibles on parle d'informations de santé protégées aux États-Unis, qui regroupent des catégories assez similaires à celles sous le coup du RGPD. On ne parlera pas de responsable de traitement ou de sous-traitants mais des "covered entity" pour désigner les entités qui sont directement soumises à la réglementation, telles que les fournisseurs de

⁴⁸ https://esante.gouv.fr/sites/default/files/media_entity/documents/doctrine--technique-du-numerique-en-sante_version-2019-comparee_vf.pdf

⁴⁹ <https://www.cdc.gov/nhsn/hipaa/index.html>

⁵⁰ <https://www.proofpoint.com/fr/threat-reference/hipaa-compliance>

soins de santé, les compagnies d'assurance maladie et les organismes gouvernementaux. La HIPAA ne fait pas de distinction entre les responsables du traitement et les sous-traitants de la même manière que le RGPD. La Privacy Rule, ou règle de confidentialité, est la règle principale qui régit cette loi. Il s'agit du cadre réglementaire pour la protection des informations de santé identifiables individuellement, qu'elles soient détenues ou transmises par une entité couverte ou son partenaire commercial. La réglementation énonce des directives spécifiques quant à l'utilisation et la divulgation des informations de santé protégées (PHI), stipulant entre autres, l'obligation de consentement et de notification. Par ailleurs, elle confère aux patients le droit d'accéder à leurs propres PHI, leur permettant de solliciter des copies de leurs dossiers médicaux et d'effectuer des corrections si nécessaire. Les entités couvertes sont également tenues d'implémenter des politiques et des procédures de protection des PHI, et de désigner un responsable dédié pour superviser ces efforts de mise en œuvre. On constate donc les similitudes entre les deux réglementations et les finesses dans l'approche de sécurisation de la donnée.

Un autre élément clé de l'HIPAA est la règle de sécurité, qui existe au sein de la règle de confidentialité. Ce sous-ensemble est constitué de toutes les informations de santé identifiables individuellement qu'une entité couverte crée, reçoit, conserve ou transmet sous forme électronique, les ePHI. Tout comme les obligations du RGPD, la conformité à la Security Rule de la HIPAA doit être considérée comme une culture à part entière, intégrée et vivante au sein des établissements de santé pour assurer la protection efficace de la vie privée, de la sécurité et de l'intégrité des ePHI.

La pertinence d'évoquer cette législation ici réside dans l'apparente similarité avec la législation européenne mais qui contient des écueils. Cela peut mener les pays voulant avoir une politique de donnée de santé véritablement souveraine à éprouver certaines difficultés en choisissant un fournisseur provenant des États-Unis. Ainsi, toute entreprise là-bas est soumise au Patriot Act, signé en 2001, et qui prévoit très grossièrement que les agences gouvernementales ont le pouvoir de récupérer, auprès des opérateurs de télécommunication, toutes les informations personnelles d'un usager sur simple soupçon, et sans qu'il en soit informé.⁵¹ Cela inclut bien entendu les données de santé et ne se restreint

⁵¹ <https://www.congress.gov/107/plaws/publ56/PLAW-107publ56.pdf>

pas aux citoyens américains soupçonnés de collusion avec le terrorisme.⁵² Les affaires liées à la NSA et la surveillance de masse, véritables scandales géopolitiques, parachèvent une méfiance envers les solutions outre-Atlantique.

En France, notamment, la souveraineté du Health Data Hub est soulevée dans le débat public. En effet, la solution choisie jusqu'en 2025 est Microsoft Azure, soumise au Patriot Act et donc n'assurant pas la certitude de la souveraineté de la donnée.

III. Application

1. Identification des lacunes du système de santé (cybersécurité, perte de données, structure, interopérabilité)

Le système de santé français a été un gage d'excellence et un métronome pour le monde, mais perd aujourd'hui de sa superbe. En outre, un rapport de l'OMS en 2000 nous classait, à l'époque, premiers des systèmes de santé mondiaux.⁵³ Cette place s'est considérablement dégradée depuis, entre une 20^{ème} place dans le Legatum Prosperity Index 2023 et une sortie du top 10 des pays qui offrent les meilleurs soins de santé en Europe, selon l'indice Euro Health Consumer Index/Health Consumer Powerhouse (HCP).⁵⁴ Entre autres problèmes, le nombre de lits disponibles en diminution et des services d'urgences débordés, peuvent parfois conduire à des décès inattendus de patients du fait de l'attente trop longue.⁵⁵ Sujet principal ces derniers temps, de nombreuses cyberattaques viennent noircir un tableau clinique qui se dégrade.

⁵² . https://www.lemonde.fr/les-decodeurs/article/2015/01/12/le-patriot-act-une-legislation-d-exception-au-bilan-tres-mitige_4554570_4355770.html

⁵³ https://www.medcost.fr/html/economie_sante_eco/eco_070700.htm.

⁵⁴ https://www.prosperity.com/download_file/view_inline/4789

⁵⁵ <https://www.hospimedia.fr/actualite/analyses/20230223-gestion-des-risques-l-augmentation-des-deces-inattendus>

4.1 Chiffres clés pour la période 2021-2022



Figure issue du rapport cité en ⁵⁶

On estime que la France subit environ une attaque d'envergure par semaine sur un établissement de santé, ce qui est colossal et insoutenable sur le long terme. Ce n'est d'ailleurs pas le seul pays et l'on se retrouve face à un fléau international. Des groupes de cybercriminels, en particulier russes comme la Vice Society, revendiquent de nombreuses actions offensives sur des établissements de santé français, dans un climat géopolitique instable sur fond de guerre en Ukraine. ⁵⁷ L'impact financier et sur les pratiques est tout à fait majeur pour ces établissements qui perdent, en plus des données patients, le paramétrage de leurs dispositifs médicaux, la programmation des rendez-vous sur plusieurs mois, l'obligation de repasser sur format papier quelques temps et de restructurer le SI de l'hôpital.

Nous avons pu conduire des entretiens avec des DSI d'hôpitaux dans le cadre de mon travail, et notamment pour répondre à des demandes du gouvernement pour la constitution d'une task force Cybersécurité. L'entreprise GALEON a été choisie et a contribué avec les notes de ces entretiens, dont nous pouvons communiquer certains extraits en anonymisant les données. Il en ressort des écueils similaires dans beaucoup d'établissements de santé français. Le premier établissement, dans le Nord de la France, a apparemment atteint un

⁵⁶

https://esante.gouv.fr/sites/default/files/media_entity/documents/ans_certsante_rapport_public_observatoire_signalements_issis_2022_vf.pdf

⁵⁷ https://www.lemonde.fr/pixels/article/2022/10/19/a-paris-la-maternite-des-bluets-frappee-par-une-attaque-informatique_6146539_4408996.html

niveau remarquable d'équipement en matière de cybersécurité. En plus des pare-feu périmétriques et latéraux, le CHU est équipé de bastions, de protections antispam, d'un système de sécurisation des courriels via certains protocoles, et d'une segmentation de réseau. En revanche, dans les établissements du Groupe Hospitalier Territorial (GHT) qui lui sont affectés, la situation est moins reluisante. La moitié des responsables informatiques du GHT peinent à gérer le firewall, et 80% des établissements du GHT, généralement très petits avec un seul technicien sur place, dépendent de l'établissement mère pour le contrôle du firewall. La stratégie du GHT consiste à centraliser les outils de sécurité et l'administration au niveau du support, en maintenant un socle incompressible. Actuellement, chaque établissement héberge son propre Système d'Information. L'établissement, à lui seul, gère 289 applications, 1 000 serveurs, des centaines de bornes wifi, et des éléments actifs de réseau, avec une équipe dédiée de 20-25 responsables d'applications et 12 personnes en charge de l'infrastructure.

Côté cyberdéfense, ce centre de soins effectue des audits réguliers et dispose d'une équipe de sécurité dédiée. Néanmoins, il est crucial de soutenir plusieurs axes prioritaires. En plus des audits nécessaires pour établir un cadre, des moyens récurrents sont nécessaires, avec des contreparties liées au financement. Il est impératif d'établir un plan d'action et des objectifs pour les petits établissements, gérés par le support. Il faut aussi allouer des moyens de façon centralisée au support avec des objectifs mesurés. L'accent doit être mis sur le financement du temps humain (ETP) et non uniquement sur l'investissement sur des équipements.

Un autre établissement, également support de son Groupe Hospitalier Territorial, se confronte à de sérieux défis en matière de cybersécurité. Le recrutement d'ingénieurs de sécurité est un problème, en raison du manque de candidats et d'attractivité financière. Un plan de sécurisation sur trois ans a été établi dès 2023, avec un budget estimé de 400k€ par an et 2.5 ETP par an. Cependant, l'établissement peine à investir la moitié de ce budget, malgré les efforts de financement par France Relance. Le CH, un site pivot avec huit établissements de proximité sous sa responsabilité, doit également composer avec ces établissements qui ne peuvent pas financer leurs propres actions de cybersécurité, en raison d'un manque de moyens et de personnel.

Avec une marge de zéro, un déficit de 10M€, des difficultés à emprunter pour investir, et des taux d'emprunts à 5%, l'établissement manque de capacités d'autofinancement. Dans ces conditions, il est impossible de respecter le plan de charge à trois ans en termes de moyens financiers et humains. Les sauvegardes sont actuellement effectuées uniquement pour l'établissement pivot, et non pour les autres établissements. On compte deux data centers actifs/passifs situés dans différentes ailes du bâtiment, hébergeant 200 logiciels sur des serveurs différents et avec des versions différentes.

Au niveau de la cyberdéfense, un test d'intrusion a été réalisé et a abouti à une note de D+ sur une échelle de A à E. L'objectif est d'atteindre B+ dans trois ans. Le dossier pour la norme ISO 27001 nécessite des ressources internes importantes.

Il est donc essentiel de soutenir plusieurs axes prioritaires, notamment sanctuariser les budgets dédiés à la sécurité, attribuer un pourcentage du budget total de l'établissement à la sécurité, étendre la sauvegarde des huit petits établissements sur le pivot, et investir dans un autre data center, l'authentification par carte CPS, la centralisation et l'agrégation des logs, la formation anti-phishing, un référent numérique soignant par pôle, et la convergence des outils métiers pour faciliter le MCO.

Dans un autre GHT, malgré le financement de certains projets de cybersécurité, le manque de ressources humaines locales empêche leur déploiement. Le GHT n'a pas les moyens de recruter un ingénieur dédié à la cybersécurité pour l'ensemble du territoire, un problème exacerbé par des difficultés d'attractivité pour le recrutement d'experts SI/Sécurité. Cette lacune se fait sentir dans l'entretien quotidien post-installation d'un de leur projet de sécurité, l'analyse du contenu des outils de monitoring, les mises à jour sécuritaires et les patches réguliers des firewalls, difficiles à réaliser dans ces conditions.

Pour soutenir ces axes prioritaires, le GHT souligne la nécessité de personnel dédié, d'allouer des financements du PMSI à la DSI, de privilégier les coûts de fonctionnement sur l'investissement et d'augmenter la part du SI dans le budget de l'hôpital. De plus, le GHT propose de mettre en place une messagerie mail nationale, afin d'éviter que chaque établissement gère sa propre messagerie, et de privilégier la durabilité dans les projets, avec un horizon de 4 à 6 ans.

Un dernier établissement fait également face à des problématiques similaires. Avec un personnel informatique limité à un ETP, le DSI ne peut consacrer que 30% de son temps à cet établissement, qui ne dispose pas de RSSI. Les infrastructures informatiques sont vieillissantes, avec des serveurs datant de 2015 et une téléphonie datant de 2010. L'hôpital a subi une cyberattaque récemment, malgré des tests d'intrusion réalisés en amont. Pour soutenir cet établissement, il est recommandé d'augmenter les moyens humains pour maintenir et faire évoluer le parc informatique, d'investir en matériel de protection des SI, d'améliorer l'authentification sur les postes clients et de passer toutes les applications critiques en mode SAAS.

2. Application au domaine de la santé

Il semble pertinent de souligner que l'utilisation des données structurées en santé présente un potentiel considérable pour pallier certaines des lacunes évoquées du système de santé français. Tout d'abord, une adoption plus généralisée de données structurées faciliterait l'interopérabilité entre les différents acteurs du système de santé, tels que les hôpitaux, les médecins, les laboratoires et les pharmacies. Cela permettrait une communication plus fluide et efficace, améliorant ainsi la coordination des soins et la prise en charge des patients. En Estonie, le taux de digitalisation des données de santé atteint 95% alors qu'en France il n'est pas rare de voir des services entiers reposer sur des dossiers papier. Près de 100 % des résidents estoniens ont un dossier de santé numérique, presque l'intégralité des ordonnances sont numérisées et un service de consultation en ligne facilite la collaboration entre les médecins traitants et les médecins spécialisés. Une seule caisse d'assurance maladie couvre 95% de la population pour 102 CPAM en France.⁵⁸

La décentralisation des données de santé veut aussi dire qu'en cas de cyberattaque sur un établissement, le patient aura beaucoup plus de chances d'avoir ses données répliquées sur un autre serveur et donc disponibles tout de même. L'utilisation du format papier présente certes certains avantages, notamment l'habitude pour les anciennes générations et la bonne idée de ne pas être piratable à distance.

⁵⁸ <https://www.france-estonie.org/estonie-un-exemple-europeen-detat-numerique-performant/>

Le Service Européen de Santé en Ligne, ou Sesali, est une initiative coordonnée par l'Agence Numérique en Santé (ANS) sous l'autorité du Ministère chargé de la santé, en partenariat avec la Commission européenne et une vingtaine de pays membres de l'Union européenne.

⁵⁹ Cette plateforme numérique vise à faciliter l'accessibilité et la compréhension transfrontalière des synthèses médicales des patients, en levant les obstacles linguistiques. Des informations pertinentes comme les allergies, les traitements actuels, les antécédents pathologiques et chirurgicaux peuvent être traduites et affichées en français pour les professionnels de santé autorisés en France.

Dans l'objectif d'assurer la continuité des soins à l'échelle de l'UE, Sesali a le potentiel d'être enrichi progressivement pour couvrir un éventail plus large de services, bénéficiant à l'ensemble de la population européenne.

En ce qui concerne la gestion des données des patients, le système Sesali garantit un accès restreint aux données de santé des patients européens, exclusivement pour les professionnels de santé autorisés en France, directement impliqués dans les soins aux patients, et assujettis à la confidentialité professionnelle. Le consentement du patient est nécessaire et les données ne seront pas réutilisées pour des objectifs autres que ceux liés à la prise en charge sanitaire.



⁵⁹ <https://esante.gouv.fr/produits-services/sesali>

Ensuite, les données structurées pourraient contribuer à renforcer la recherche médicale et l'innovation en France. En rendant les données plus facilement accessibles et exploitables, les chercheurs pourraient analyser de vastes ensembles de données pour identifier des tendances, des corrélations et des facteurs de risque, ce qui faciliterait le développement de nouvelles thérapies, de diagnostics améliorés et de stratégies de prévention.⁶⁰

Exploiter au mieux les données structurées permettrait une meilleure évaluation de la qualité des soins et de la performance des professionnels de santé. Les autorités sanitaires et les établissements de soins peuvent déterminer des indicateurs clés de performance (KPI) qui reflètent la qualité et l'efficacité des soins prodigués. Ces indicateurs peuvent inclure des mesures telles que les taux de réadmission, les complications post-opératoires, les infections nosocomiales, la durée des séjours hospitaliers et les délais d'attente pour les consultations par exemple. La collecte des données nécessaires pour passer certaines certifications imposées aux établissements de santé, par la HAS par exemple, pourrait être accessible par l'export d'un fichier à partir d'une requête sur une base de données structurées. Les autorités de santé pourraient suivre les indicateurs clés de performance et mettre en œuvre des mesures correctives en cas de problèmes identifiés, contribuant ainsi à l'amélioration continue des services de santé.

Structurer les données pourrait également aider à optimiser la gestion des ressources et la planification des soins au sein du système de santé, que ce soit à l'échelle d'un établissement ou du territoire. Grâce à l'analyse sur un nombre considérable de données et en flux tendu, les décideurs pourraient mieux anticiper les besoins en matière de soins, d'équipements et de personnel, et ainsi allouer les ressources médicales, les lits de manière plus efficace et équitable.

L'analyse des données structurées permet également d'identifier les variations dans les pratiques médicales et les écarts par rapport aux recommandations et aux directives établies. En repérant les zones d'amélioration, les professionnels de santé peuvent se concentrer sur la mise en œuvre de pratiques fondées sur des preuves scientifiques et ajuster leurs approches en fonction des besoins spécifiques de leurs patients. On peut

⁶⁰ https://esante.gouv.fr/sites/default/files/media_entity/documents/sesali_guide_utilisateur.pdf

accéder, en plus d'une science basée sur les preuves, à une science basée sur un nombre massif de données, qui correctement collectées et conçues constitueraient un niveau de preuve supplémentaire.

Une analyse en plus haute instance de ces données pourrait conduire à l'établissement de comparaisons entre les établissements de santé et les praticiens, en mettant en évidence les domaines où la performance est supérieure ou inférieure à la moyenne. Cela encourage une émulation positive et incite les professionnels à adopter les meilleures pratiques identifiées chez leurs pairs. D'un autre côté, il ne faut pas créer un climat de compétition et de délation qui augmenterait le stress des praticiens en punissant les établissements en deca de la moyenne.

IV. Perspectives

1. Revue des projets existants et futurs sur la donnée structurée en santé.

Outre les intégrations de la SNOMED-CT et de la CIM-11, que nous avons vues plus haut, les projets autour de la donnée structurée sont légion. L'échange d'informations dans le système informatique est une tâche difficile, et de nombreux acteurs internationaux et nationaux s'attèlent à essayer de standardiser les méthodes. L'un des meilleurs exemples de cette volonté est le Fast Healthcare Interoperability Resources (FHIR). FHIR est une norme développée par Health Level Seven International (HL7), organisation à but non lucratif dédiée au développement de l'interopérabilité des données de santé et la standardisation du protocole d'échanges médicaux.⁶¹ Le L7 de HL7 indique que c'est un langage qui fonctionne au niveau de la couche 7, c'est-à-dire la couche application, du modèle OSI.⁶²

Aux États-Unis, l'Observational Health Data Sciences and Informatics (OHDSI), que l'on prononce « Odyssey », est une collaboration internationale qui vise à transformer les données de santé en connaissances exploitables.⁶³ Le projet développe des outils et des méthodes pour harmoniser et analyser de manière structurée les données de santé provenant de différentes sources, afin de soutenir la recherche clinique et la prise de

⁶¹ <https://fhir.org/>

⁶² . <https://www.cloudflare.com/fr-fr/learning/ddos/glossary/open-systems-interconnection-model-osi/>

⁶³ <https://www.ohdsi.org/>

décision en matière de soins. Ces outils sont souvent open source et ils proposent des formations.

Du côté de la génétique, qui est le sujet scientifique qui a uni le plus fort la communauté scientifique internationale, les initiatives ne sont pas en reste. On se souvient du séquençage quasi-complet du génome humain en 2003. Créée en 2013, Global Alliance for Genomics and Health (GA4GH) ⁶⁴ a pour mission de faire progresser la santé humaine grâce aux données génomiques. Elle se compose de plus de 500 organisations dans divers domaines, s'efforçant de développer des normes mondiales à travers un processus pré-concurrentiel et inclusif. Le but de la GA4GH est de réaliser le potentiel de la médecine génomique en facilitant l'accès aux informations génétiques pour les chercheurs et médecins. Cependant, des obstacles subsistent, notamment le stockage des données génomiques dans des systèmes incompatibles. Afin de surmonter ces défis, la GA4GH vise à construire un "internet de la génomique", assurant à la fois la protection des données personnelles et le partage de connaissances vitales. Pour ce faire, elle développe des produits libres et open source, facilitant la gestion sécurisée et efficace des données génomiques. Par exemple, une des créations de la GA4GH est l'Human Phenotype Ontology. C'est un projet qui vise à développer une ontologie standardisée pour décrire les phénotypes humains. Il facilite l'analyse structurée des données phénotypiques pour la recherche génétique, la médecine de précision et le diagnostic des maladies rares.

On retrouve dans tous ces projets des valeurs communes qui prennent racine dans leurs origines, ancrées dans la culture internet. Une volonté de simplification, d'universalité, une tendance à l'open source et la mise en commun des ressources se retrouvent à chaque fois. Cela peut parfois être au profit de la population d'un pays seulement, ou des utilisateurs d'un logiciel, mais la mondialisation et les échanges d'informations font que ces initiatives finissent par infuser à une plus grande échelle.

⁶⁴ <https://www.ga4gh.org/>.

a. Le SNDS et le HDH

Créé en 2016, L'objectif du Système National de Santé est de faire de la France l'un des pays pionniers dans le domaine de la promotion et de la valorisation des données de santé. Géré par la Plateforme des données de santé (PDS ou « Health data hub ») et la Caisse nationale de l'assurance maladie (CNAM), le SNDS est alimenté par différentes bases de données nationales, notamment le « SNIIRAM » (Système national d'information inter-régimes de l'assurance maladie) contenant les données relatives à toutes les dépenses de l'assurance maladie, la base de données « PMSI » (Programme de médicalisation des systèmes d'information) contenant les données d'analyse de l'activité des établissements de santé et la base de données du CépiDc (Centre d'épidémiologie sur les causes médicales de décès) contenant les données relatives aux causes de décès. les données relatives au handicap ; un échantillon de données en provenance des organismes d'Assurance Maladie complémentaire.⁶⁵

Le projet national de structuration des données de santé en France : Ce projet vise à créer un cadre national pour la structuration, l'échange et l'utilisation des données de santé, en mettant l'accent sur l'interopérabilité et la conformité réglementaire. Il implique la mise en œuvre de standards tels que la CIM-11, la SNOMED CT et le dossier médical partagé (DMP).

Pour gérer cette évolution majeure, en 2016, les autorités ont mis en place le Système national des données de santé (SNDS), qui compile et met à disposition des informations de santé "pseudonymisées" recueillies par des organismes publics.

Le Health Data Hub quant à lui a été créé par une loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé. Le HDH associe 56 acteurs en grande partie publics, comme le CNAM, CNRS, Haute Autorité de santé, France Assos Santé...etc. C'est une sorte de prolongement du SNDS, qui permet à des particuliers porteurs de projets, en général des chercheurs, d'effectuer des demandes d'accès à la base de données.

Le Health Data Hub est une structure qui est reliée à plusieurs polémiques actuellement, notamment du fait de l'hébergement chez Microsoft Azure que nous avons déjà évoqué,

⁶⁵ <https://www.snds.gouv.fr/SNDS/Accueil>

acteur américain par conséquent susceptible de tomber sous la coupe du Patriot Act. Le député Philippe Latombe a ainsi saisi la Commission d'accès aux documents administratifs (Cada) un rapport jamais publié par le HDH⁶⁶. En effet, depuis 2019, le Hub justifie son recours à Azure par une infériorité technique et de sécurité des hébergeurs français et européens, s'appuyant sur ce rapport privé. Il faut savoir que l'ex-ministre de la Santé Olivier Véran avait donné au HDH jusqu'en novembre 2022 pour changer de solution, mais le Hub a repoussé, en décision interne, l'échéance à 2025.

Ces projets et d'autres initiatives similaires illustrent l'importance croissante de la donnée structurée en santé. Ils témoignent de la politique française de développer et coordonner la collaboration entre les acteurs du domaine de la santé pour développer des normes, des outils et des méthodes qui favorisent l'interopérabilité, la recherche et l'amélioration de la qualité des soins⁶⁷.

b. Les données structurées dans la pandémie :

En réponse à la pandémie de COVID-19, la France a mis en œuvre plusieurs stratégies axées sur la structuration des données. Nous nous souvenons du Système d'Informations de DEPistage (SI-DEP), un outil conçu par Santé publique France pour rationaliser la gestion et le suivi des tests COVID-19.⁶⁸ Cette plateforme compile les données des tests effectués, y compris les résultats des tests PCR et antigéniques, ainsi que des informations pertinentes sur les individus testés. Ces données ont servi et continuent de servir à surveiller l'évolution de la pandémie, à identifier les foyers de contagion et à mettre en place des mesures de contrôle appropriées.

Une autre initiative significative est l'application TousAntiCovid, déployée par le gouvernement le 22 octobre 2020, en vue de soutenir les efforts de traçage des contacts dans le contexte de la pandémie de COVID-19. Cette application, présente sur plus de 62 millions de téléphones en France, était une version améliorée de l'application StopCovid, enrichie par l'inclusion d'informations factuelles et sanitaires sur l'épidémie. Son objectif

⁶⁶ <https://www.apmnews.com:443/story.php?objet=398916>

⁶⁷ <https://www.healthandtech.eu/fr/article/view/16741/experimentations-article-51-role-cle-donnees-snds-evaluations-analyse.html>

⁶⁸ <https://sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/coronavirus/professionnels-de-sante/article/enregistrement-des-resultats-de-tests-antigeniques-dans-si-dep>

principal était de permettre aux utilisateurs d'être alertés ou d'alerter d'autres personnes en cas d'exposition au COVID-19, contribuant ainsi à briser les chaînes de transmission et à ralentir la propagation du virus. Cependant, avec l'amélioration de la situation épidémiologique, comme stipulé par la loi du 30 juillet 2022, les dispositifs de traçage des contacts ont été désactivés le 25 janvier 2023. TousAntiCovid continue de fournir un carnet pour stocker les certificats de vaccination et de test, qui peuvent s'avérer utiles lors des voyages ou pour obtenir un remboursement de test. De plus, elle maintient l'accès à des statistiques clés et des actualités pour tenir les utilisateurs informés de la situation.

Ces initiatives démontrent clairement le rôle crucial des données dans la gestion de la pandémie de COVID-19 en France. Grâce à ces efforts, il a été possible de suivre plus efficacement l'évolution de l'épidémie, d'optimiser les approches de dépistage et de vaccination, et d'améliorer les soins aux patients atteints de COVID-19.

2. L'intérêt pour la recherche médicale

La recherche médicale est un milieu difficile et exigeant qui nécessite de produire beaucoup de publications pour progresser dans les échelons et être bien considéré de ses pairs. Le stress occasionné peut parfois conduire les chercheurs à altérer la qualité, voire la nature, des résultats et leur reproductibilité.⁶⁹ On parlera même de véritable « crise de la reproductibilité », terme venu d'un article publié dans Nature en 2016 où 70% des chercheurs admettaient ne pas être capables de reproduire les résultats de recherches provenant d'autres équipes.⁷⁰ De plus, une culture du « Publish or Perish » s'est de plus en plus instaurée dans le monde scientifique, les chercheurs devant produire un maximum de contenu pour atteindre certains indicateurs chiffrés et être acceptés dans des revues à plus haut facteur d'impact. Il en découle une baisse de qualité globale au niveau des publications et des données utilisées. Disposer de données structurées, facilement interprétables et répondant à des normes internationales peut améliorer la collaboration entre chercheurs. La publication des données brutes de recherche en fin de publication n'est pas une pratique courante, mais elle est de plus en plus encouragée dans certains domaines, comme sur le

⁶⁹ <https://doi.org/10.1038/533452a>

⁷⁰ https://www.lemonde.fr/sciences/article/2018/10/23/la-biologie-francaise-minee-par-l-inconduite-scientifique_5373162_1650684.html

site <https://arxiv.org/>⁷¹ dont sont issues certaines sources de cette thèse et qui référence plus de deux millions d'articles open source. Améliorer la qualité de la donnée permet d'accélérer le partage de connaissances. De plus, publier les données et les rendre disponible peut améliorer la reproductibilité et encourager d'autres équipes de recherche à les utiliser dans d'autres contextes. Dans le cadre d'études, ces données sont pseudonymisées, c'est-à-dire qu'on peut retracer leur origine si cela est nécessaire, ou anonymisées totalement, donc à la réidentification impossible. La réalisation de méta-études en serait elle-même beaucoup plus simples, puisque les données suivraient un référentiel standardisé plus comparable. Toute la recherche comparative pourrait en bénéficier.

Les données structurées permettent d'augmenter la cohérence, la précision et la fiabilité des données utilisées. Avoir une meilleure connaissance de la donnée que l'on collecte et manipule permet de réduire considérablement les biais qui la concernent.

Disposer d'une base de données rangée, incrémentale en temps réel et requêtable à tout moment peut également être un énorme atout pour les études sur les tendances épidémiologiques. Cela pourrait être exploité à l'échelle de toute la population d'un pays si l'infrastructure nécessaire est mise en place.

Nous pouvons ainsi citer une étude qui utilise le deep learning pour découvrir et classifier trois sous-phénotypes d'insuffisance rénale aiguë en utilisant des données structurées et non structurées de dossiers de santé issus de DPI de patients, avant le diagnostic d'insuffisance rénale aiguë⁷². Ces sous-phénotypes sont caractérisés par des degrés variables de perte de la fonction excrétrice rénale et des biomarqueurs spécifiques. Cet exemple d'alliance de données structurées et non structurées issus de DPI est un exemple de la puissance d'avoir énormément de données à disposition et harmonisées. Connaître plus en détail ces trois phénotypes va ainsi améliorer la compréhension de la physiopathologie de la maladie et le développement d'interventions cliniques plus ciblées.

Les données de vie réelle ou « real world data » sont des données qui n'interviennent pas sur la prise en charge des patients et ne sont pas collectées dans un cadre expérimental hospitalier. Au lieu de cela, elles sont générées et rassemblées durant les soins habituels ou

⁷¹ <https://arxiv.org/>

⁷² <https://doi.org/10.1016/j.jbi.2019.103361>

par le biais d'appareils connectés. Les données de vie réelle représentent en principe les pratiques habituelles. Le potentiel de ces données de santé de la vie quotidienne a été accentué récemment. Leur utilisation facilite l'acquisition de nouvelles informations, en améliorant la capacité de suivre l'utilisation et l'efficacité des produits de santé, tout en accélérant la recherche. Elles diffèrent des données cliniques utilisées pour obtenir une autorisation de mise sur le marché (AMM). En effet, les demandes de données de vie réelle post-AMM auprès des autorités réglementaires sont de plus en plus nombreuses. Différents acteurs produisent ces données spécifiques : les laboratoires pharmaceutiques, les CRO, les médecins et les patients (notamment grâce aux appareils connectés). Les industriels lancent des études spéciales et se basent sur de grandes bases de données existantes, comme le SNDS, pour effectuer notamment des études médico-économiques. Les sources de ces données spécifiques sont diverses (dossiers de patients, applications mobiles, solutions connectées, réseaux sociaux) tout comme leurs analyses et traitements via des agrégateurs de données, tels que le Health Data Hub, ou des entrepôts de données, qu'ils soient hospitaliers ou non. Il est tout à fait possible de structurer également les données de vie réelle et en tirer tous les avantages qui en découlent.⁷³

3. Intérêt pour les algorithmes d'IA (Machine learning, deep learning)

Les caractéristiques des données structurées que nous avons citées jusqu'à présent ont une valeur très importante pour les algorithmes d'intelligence artificielle. Le terme d'IA défraie la chronique en ce moment et il possède une définition multiple. Une vague impressionnante de nouveaux outils se multiplie, notamment grâce à l'avènement des LLMs, ou Large Language Models. Ce sont de grands modèles de langage, des réseaux de neurones profonds entraînés sur une quantité astronomique de texte et avec un grand nombre de paramètres. L'ambition est d'assister l'être humain sur des tâches complexes, faciliter les recherches, et parfois en santé, aider au diagnostic. Comme dit plus haut, on entend « intelligence » dans IA au sens de renseignement, c'est-à-dire que l'agent est capable de chercher de l'information dans un corpus prédéfini, mais pas d'innover véritablement. Cependant, cette capacité de recherche peut mener à des réponses qui peuvent surprendre ou aider un utilisateur humain en sortie. Nous ne détaillerons pas la technologie ici, mais l'IA est de plus

⁷³ <https://www.healthandtech.eu/fr/article/view/16247/vers-bases-donnees-entierement-geneeres-ordinateur-etudes-rwe-sans-patients.html>

en plus utilisée sur des données médicales, et ce depuis plus longtemps que la récente résurgence d'intérêt. Les avancées récentes ont vu la prolifération de chatbots destinés aux patients pour les accompagner dans la maladie et améliorer l'observance, ou d'algorithmes entraînés sur un large nombre de photographies de tumeurs potentielles et capable de reconnaître lesquelles en sont véritablement.⁷⁴ Des avancées majeures dans le repliement des protéines sont également devenues possibles⁷⁵. Bien que l'aide à la décision soit utile à un soignant, il ne faut en aucun cas que ces algorithmes remplacent l'intégralité du processus diagnostique. La loi en France qui concerne les décisions algorithmiques en santé est l'article 17 de la loi n° 2021-1017 du 2 août 2021, également connue sous le nom de "nouvelle loi de bioéthique".⁷⁶ Cet article vise à encadrer le traitement algorithmique des données massives et les décisions médicales qui en découlent. Il stipule que toute personne a le droit de ne pas faire l'objet d'une décision individuelle prise exclusivement sur la base d'un traitement algorithmique.

Que ce soit en deep learning ou en machine learning, les algorithmes ont besoin de données d'entrée de qualité pour fournir des résultats pertinents. Les apports concrets de la donnée structurée dans ce sens seraient d'éviter le phénomène de GIGO (Garbage In, Garbage Out). En effet, une donnée de piètre qualité ne pourra jamais donner des résultats faramineux. Les biais inhérents à la donnée peuvent être mieux anticipés, et les résultats peuvent être mieux expliqués. Il existe un problème de transparence dans le fonctionnement des algorithmes et la méthode de raisonnement employée.⁷⁷ Utiliser des données de meilleure facture pourrait également faciliter et améliorer la performance des algorithmes, tout en assurant un respect de la réglementation. De nombreuses institutions publiques et privées réfléchissent à l'éthique derrière l'utilisation de l'IA en santé, qui sera un enjeu majeur du 21^{ème} siècle.⁷⁸

79

⁷⁴ <https://doi.org/10.1038/s41598-022-22514-4>

⁷⁵ <https://directorsblog.nih.gov/2021/07/27/artificial-intelligence-accurately-predicts-protein-folding/>.

⁷⁶ <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043884384/>

⁷⁷ <https://www.cairn.info/revue-francaise-d-administration-publique-2018-3-page-525.htm>

⁷⁸ <https://Www.Consilium.Europa.Eu/Fr/Press/Press-Releases/2022/12/06/Artificial-Intelligence-Act-Council-Calls-for-Promoting-Safe-Ai-That-Respects-Fundamental-Rights/>

⁷⁹ <https://www.who.int/fr/publications-detail/9789240037403>

4. Défis des données structurées

Nous avons pu voir de nombreux avantages dans la structuration plus systématique des données, mais il est bon de rappeler que ce n'est pas la panacée. En effet, des problèmes très concrets peuvent freiner l'adoption, on même justifier de ne pas adopter la structuration des données tout court. Les schémas et modèles préétablis pour la conception ou le stockage des données peuvent être considérés comme trop peu flexibles si l'on veut représenter des informations complexes et changeantes. Dans ces cas-là, on privilégiera les données non structurées ou semi-structurées, en complément de données structurées ou non. Une autre problématique qui émane de ça est le risque de perte d'information, car structurer à outrance des champs de données peut affecter la quantité et la qualité d'un commentaire d'un soignant par exemple, qui ne pourra pas communiquer sur un aspect d'une maladie ou de l'état d'un patient non prévus par le logiciel, s'il n'y pas de champ de texte libre à côté des champs structurés par exemple. C'est une question d'équilibre entre structuration des données et préservation de la richesse d'informations à long terme. Maintenir une infrastructure robuste a également un coût en temps, efforts et ressources qui peut ne pas être négligeable pour les établissements. Il est nécessaire de disposer des compétences et sensibiliser en interne sur l'importance de collecter et organiser les données selon certaines normes. En ce sens, un haut niveau de complexité au niveau des données peut parfois rendre le respect du RGPD et de la HIPAA plus difficile, particulièrement si les normes mises en place sont nouvelles et innovantes.

Posséder une avance en termes de normes peut-être à double tranchants. Des systèmes moins novateurs qui utilisent des normes d'interopérabilité antérieures peuvent alors devenir hermétique à la communication. C'est souvent une des raisons de l'inertie des systèmes informatiques hospitaliers, en plus du manque de moyens financiers. Certains établissements ont des profils réputés plus réfractaires à l'innovation et peuvent ne pas prendre le risque de changer tout un système moins performant mais qui a fait ses preuves, ce qui est tout à fait understandable. L'interopérabilité restant un milieu commercial, il y a parfois de la compétition et il peut être difficile d'exiger que les applications de deux concurrents communiquent parfaitement.

La transition vers l'adoption plus généralisée des données structurées représente un défi majeur pour les professionnels de santé, car elle nécessite un changement culturel et

organisationnel profond. Les soignants doivent non seulement apprendre à maîtriser de nouveaux outils et technologies, mais aussi repenser leurs processus de travail et assumer de nouvelles responsabilités. L'adoption de données structurées implique la mise en place de systèmes d'information interconnectés et la formation continue des professionnels de santé pour garantir une utilisation efficace et sécurisée des données. La collaboration entre les équipes médicales, informatiques et administratives devient essentielle pour assurer la cohérence et la qualité des données collectées et partagées. La transition réussie vers des systèmes de données structurées dépend de la capacité des professionnels de la santé à s'adapter et à embrasser ces changements, reconnaissant le potentiel qu'ils offrent pour améliorer les soins aux patients, optimiser les ressources et faciliter la recherche médicale.

5. Intérêt pour les essais in silico

La structure traditionnelle de la recherche clinique nécessite des processus extrêmement longs et coûteux, en termes de matériel et de ressources humaines. L'un des domaines les plus intéressants économiquement et qui s'est le plus développé ces dernières années est certainement celui des essais in silico, c'est-à-dire simulés par calculs informatiques. On parle de « in silico » en référence au Silicium souvent utilisé dans les composants électroniques⁸⁰. En outre, la pandémie de COVID-19 a été un moteur inopiné de cette digitalisation, en mettant en évidence la nécessité de nouvelles approches pour mener des essais cliniques pour la sécurité de tous les protagonistes. Cette propension est accélérée par la démocratisation d'outils informatiques du côté des promoteurs et investigateurs de la recherche. Ces simulations nécessitent des données au plus proche de la réalité clinique.

Les données structurées jouent un rôle essentiel dans l'amélioration des essais in silico, qui sont des simulations informatiques utilisées pour modéliser et analyser les processus biologiques, les interactions médicamenteuses et les réponses thérapeutiques.⁸¹ Grâce à la structuration des données, il est plus facile de collecter, gérer et intégrer des informations provenant de diverses sources, telles que les données génomiques, protéomiques, métabolomiques et cliniques. Ces données peuvent ensuite être utilisées pour alimenter et valider des modèles mathématiques et des simulations informatiques.

⁸⁰ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1978274/>

⁸¹ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8979333>

En exploitant des données structurées de haute qualité, les essais in silico peuvent gagner en précision, en fiabilité et en robustesse. Cela permet aux chercheurs de mieux comprendre les mécanismes biologiques sous-jacents, de prédire les effets des interventions thérapeutiques et d'identifier les cibles médicamenteuses potentielles. De plus, les données structurées facilitent l'interopérabilité entre les différentes plateformes et outils informatiques utilisés dans les essais in silico, favorisant ainsi la collaboration entre les chercheurs et la réplique des résultats. (74)

Les essais in silico bénéficient également de l'exploitation des données structurées pour accélérer le processus de découverte et de développement des médicaments. En combinant des simulations informatiques avec des données structurées provenant d'études précliniques et cliniques, les chercheurs peuvent optimiser la sélection des candidats médicaments, réduire les risques associés au développement et minimiser les coûts et les délais. Les essais in silico peuvent également contribuer à l'avènement de la médecine personnalisée, en utilisant des données structurées pour modéliser les réponses individuelles aux traitements et ainsi concevoir des thérapies adaptées aux besoins spécifiques de chaque patient.

6. Intérêt pour les jumeaux numériques

L'un des sujets de santé et innovation qui a le plus défrayé la chronique ces dernières années est l'ambition de créer des répliques identiques du corps humain de manière digitale, à la manière des avatars vidéoludiques, jusque dans les moindres mécanismes biologiques. Les jumeaux numériques existent déjà dans de nombreux domaines industriels, car il est plus facile de simuler l'aérodynamisme d'une aile d'avion que l'ensemble du microbiote humain⁸². Répliquer numériquement le corps nécessite une compréhension plus que parfaite du corps humain dans son ensemble et de la biologie sous-jacente, de même qu'une puissance de calcul énorme. L'une des perspectives à long terme serait de pouvoir modéliser chaque patient numériquement et ainsi comprendre les dysfonctionnements qui lui sont propres, une forme de médecine personnalisée très poussée. Cela pourrait permettre d'adapter une intervention chirurgicale complexe à la physiologie exacte du patient ou produire une spécialité thérapeutique contenant exactement les principes actifs nécessaires à l'action

⁸² <https://doi.org/10.1109/MNET.118.2200071>

pharmaceutique souhaitée. C'est un sujet qui rejoint énormément les essais in silico dans l'utilisation des nouvelles technologies pour améliorer l'efficacité des essais cliniques, la différence résidant dans le fait que le jumeau numérique se situe plutôt au niveau de l'individu. Les jumeaux numériques en santé, peuvent être utilisés comme outils de formation pour les professionnels de santé. Les simulations et les scénarios virtuels permettent aux étudiants et aux praticiens de développer et de perfectionner leurs compétences sans risquer la santé des patients réels.

Les essais virtuels ne permettent toutefois pas encore de valider un nouveau médicament. Pour l'instant, ils ne constituent qu'une partie complémentaire des examens cliniques en laboratoire et sur les patients. Les données structurées sont un élément clé pour les jumeaux numériques en santé, qui ont besoin de données précises, fiables, correctement étiquetées et interopérables. Ces simulations peuvent aider les médecins à évaluer les risques et les bénéfices de diverses options thérapeutiques et à choisir le traitement le plus approprié ^{83 84}.

V. Conclusion

Résumé des résultats de la thèse et perspectives

Le principal objectif de cette thèse était de montrer pourquoi je pense que l'on ne peut appeler une donnée « structurée » qu'à partir du moment où l'on dispose d'une infrastructure technique et réglementaire suffisante, en particulier en France. Il existe des lacs de données à l'abandon avec des jeux de données très structurées qui ne sont pas entretenus et je pense que cela empêche sur leur degré de structuration véritable à long terme. Disposer des équipements et du personnel capables de valoriser cette donnée nécessite une formation continue et une infusion lente d'une véritable culture de la donnée dans les centres de soins. Ces derniers manquent également souvent de moyens financiers suffisants pour la mise en place de solutions qui, pour la plupart, existent déjà. En plus de cela, le paysage réglementaire autour des données change très rapidement, pour essayer de suivre le rythme effréné des innovations, notamment au niveau européen. Il est donc primordial pour toute organisation gérant des données sensibles de maintenir une veille

⁸³ <https://tulip.co/fr/blog/digital-twins-in-manufacturing/>

⁸⁴ <https://www.3ds.com/fileadmin/PRODUCTS-SERVICES/DELMIA/PDF/Whitepaper/DELMIA-APRISO-Digital-Twin-Whitepaper.pdf>

avancée de ces nouvelles règles du jeu qui peuvent modifier une activité ou un traitement de données à terme.

À travers les nombreux exemples évoqués dans cette thèse, nous avons pu voir que la donnée structurée peut impacter positivement et durablement le système de santé français et le métier de pharmacien. De très nombreux écueils seraient en partie comblés ou trouveraient un début de solution dans un traitement plus ordonné de la donnée, dans un souci global d'interopérabilité. Les données sont de plus haute qualité, plus standardisées, plus ouvertes pour améliorer la reproductibilité, grâce à une culture internet globale et mondialisée. Pour nuancer ce propos, la HAS rappelle très bien que « Les données de soins des hôpitaux ne montreront leur vrai potentiel pour les usages de la HAS qu'une fois exploitées pour des projets concrets ». ⁸⁵ Structurer des données n'est pas la panacée et ne doit pas être une condition sine qua non à leur usage pour des projets utiles à la recherche et à terme, au patient. Je voulais également montrer que les évolutions apportées font partie d'une culture et d'une mouvance qui forment un tout dans les nouvelles technologies au service de la santé en général. Il est, selon moi, important que la France continue sa politique en termes de données de santé et qu'elle continue de se placer en leader dans ce domaine. C'est une terre fertile pour l'innovation malgré les clichés, et beaucoup de professionnels de santé sont favorables à faire évoluer le système si on leur montre concrètement le bénéfice dans leur travail. Il faut cependant ne pas être naïf et rappeler que l'enjeu économique est important et que de nombreux acteurs privés subsistent à toutes les étapes du cycle de vie de la donnée, et qu'ils œuvrent eux aussi pour des standards de qualité. De nombreuses initiatives ont vu et voient encore le jour dans une offre de plus en plus étoffée et accessible. Il sera très important de prendre les virages de la CIM-11 et de la SNOMED-CT pour vraiment se distinguer comme un pays avancé en termes de standards de données.

VI. Références

Liste complète des références utilisées dans la thèse.

- (1) Statista Daily Data. « Infographie: Le Big Bang du Big Data », 19 octobre 2021.
<https://fr.statista.com/infographie/17800/big-data-evolution-volume-donnees-numeriques-genere-dans-le-monde>.

⁸⁵ https://www.has-sante.fr/jcms/p_3386076/en/entrepots-de-donnees-de-sante-hospitaliers-la-has-publie-un-panorama-inedit-en-france

- (2) Inserm. « Big data en santé · Inserm, La science pour la santé ». Consulté le 16 septembre 2023. <https://www.inserm.fr/dossier/big-data-en-sante/>.
- (3) « Attaques cybersécurité auprès d'organismes publics - uMap ». Consulté le 16 septembre 2023. https://umap.openstreetmap.fr/fr/map/attaques-cybersecurite-aupres-dorganismes-publics_821557#7/47.610/5.757.
- (4) Martignoni, Nicolas. « Qu'est-ce qu'une métadonnée », 26 septembre 2019. <https://blog.martignoni.net/2019/09/qu-est-ce-qu-une-metadonnee/>.
- (5) Google for Developers. « Bien débuter en référencement naturel (SEO) : principes de base | Google Search Central | Documentation ». Consulté le 16 septembre 2023. <https://developers.google.com/search/docs/fundamentals/seo-starter-guide?hl=fr>.
- (6) « Schema.org - Schema.org ». Consulté le 16 septembre 2023. <https://schema.org/>.
- (7) « Article L112-3 - Code de la propriété intellectuelle - Légifrance ». Consulté le 16 septembre 2023. https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006278879.
- (8) Lucidchart. « Qu'est-ce qu'un schéma de base de données ? » Consulté le 15 septembre 2023. <https://www.lucidchart.com/pages/fr/schema-base-de-donnees>.
- (9) « MySQL ». Consulté le 16 septembre 2023. <https://www.mysql.com/fr/>.
- « 1,500 scientists lift the lid on reproducibility | Nature ». Consulté le 16 septembre 2023. <https://www.nature.com/articles/533452a>.
- (10) Prévention, Ministère de la Santé et de la, et Ministère de la Santé et de la Prévention. « TousAntiCovid ». Ministère de la Santé et de la Prévention, 16 septembre 2023. <https://sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/coronavirus/article/tousanticovid>.
- (11) Glake, Daniel, Felix Kiehn, Mareike Schmidt, Fabian Panse, et Norbert Ritter. « Towards Polyglot Data Stores -- Overview and Open Research Questions ». arXiv, 12 avril 2022. <https://doi.org/10.48550/arXiv.2204.05779>.
- (12) LOI n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles (1), 2018-493 § (2018).
- (13) « Qu'est-ce ce qu'une donnée de santé ? » Consulté le 16 septembre 2023. <https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante>.
- (14) Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (s. d.). Consulté le 16 septembre 2023.
- (15) https://www.cnil.fr/sites/cnil/files/atoms/files/referentiel_-_traitements_dans_le_domaine_de_la_sante_hors_recherches.pdf
- (16) « Google - Statistics & Facts | Statista ». Consulté le 15 septembre 2023. <https://www.statista.com/topics/1001/google/>.
- (17) Drees, Jackie. « Google Receives More than 1 Billion Health Questions Every Day », 11 mars 2019. <https://www.beckershospitalreview.com/healthcare-information-technology/google-receives-more-than-1-billion-health-questions-every-day.html>.
- (18) Sheu, Yi-han, Colin Magdamo, Matthew Miller, Sudeshna Das, Deborah Blacker, et Jordan W. Smoller. « AI-Assisted Prediction of Differential Response to Antidepressant Classes Using Electronic Health Records ». *Npj Digital Medicine* 6, n° 1 (26 avril 2023): 1-12. <https://doi.org/10.1038/s41746-023-00817-8>.
- (19) Tam, Charmaine S., Janice Gullick, Aldo Saavedra, Stephen T. Vernon, Gemma A. Figtree, Clara K. Chow, Michelle Cretikos, et al. « Combining structured and unstructured data in EMRs to create clinically-defined EMR-derived cohorts ». *BMC Medical Informatics and Decision Making* 21, n° 1 (8 mars 2021): 91. <https://doi.org/10.1186/s12911-021-01441-w>.
- (20) Fong, Allan, A. Zachary Hettinger, et Raj M. Ratwani. « Exploring methods for

- identifying related patient safety events using structured and unstructured data ». *Journal of Biomedical Informatics* 58 (1 décembre 2015): 89-95. <https://doi.org/10.1016/j.jbi.2015.09.011>.
- (21) Wilkinson, Mark D., Michel Dumontier, IJbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. « The FAIR Guiding Principles for Scientific Data Management and Stewardship ». *Scientific Data* 3, n° 1 (15 mars 2016): 160018. <https://doi.org/10.1038/sdata.2016.18>.
- (22) « La science ouverte | ANR ». Consulté le 16 septembre 2023. <https://anr.fr/fr/lanr/engagements/la-science-ouverte/?mdrv=anr.fr>.
- (23) Pasteur, Ceris-Institut. « Les principes FAIR : Findable, Accessible, Interoperable, Reusable ». Open science : évolutions, enjeux et pratiques, 5 octobre 2020. <https://openscience.pasteur.fr/2020/10/05/les-principes-fair-findable-accessible-interoperable-reusable/>.
- (24) Proposition de RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL ÉTABLISSANT DES RÈGLES HARMONISÉES CONCERNANT L'INTELLIGENCE ARTIFICIELLE (LÉGISLATION SUR L'INTELLIGENCE ARTIFICIELLE) ET MODIFIANT CERTAINS ACTES LÉGISLATIFS DE L'UNION (2021). <https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A52021PC0206>.
- (25) « What Is Natural Language Processing? | IBM ». Consulté le 16 septembre 2023. <https://www.ibm.com/topics/natural-language-processing>.
- (26) « Qu'est-ce que l'annotation de données 2023 ? (Meilleurs outils, types, défis, tendances) ». Consulté le 16 septembre 2023. <https://fr.shaip.com/blog/the-a-to-z-of-data-annotation/>.
- (27) « Classification internationale des maladies (CIM) | CépiDc ». Consulté le 15 septembre) 2023. <https://www.cepiddc.inserm.fr/causes-medicales-de-deces/classification-internationale-des-maladies-cim?undefined>.
- (28) « ICD-11 2023 Release Is Here ». Consulté le 16 septembre 2023. <https://www.who.int/news/item/14-02-2023-icd-11-2023-release-is-here>.
- (29) « 11_Implementation_or_Transition_Guide_rev_25_10_2019-fr.pdf - Presearch ». Consulté le 16 septembre 2023. https://presearch.com/search?q=11_Implementation_or_Transition_Guide_rev_25_10_2019-fr.pdf.
- (30) Presearch. . « . <https://Esante.Gouv.Fr/Espace-Presse/Interoperabilite-Semantique-La-France-Choisit-La-Snomed-Ct-Pour-La-Description-Des-Localisations-Anatomiques> ». Consulté le 16 septembre 2023. <https://presearch.com>.
- (31) https://icd.who.int/fr/docs/192190_ICD-11_Implementation_or_Transition_Guide_rev_25_10_2019-fr.pdf
- (32) Fu, Mingzhou, Yu Yan, Loes M Olde Loohuis, et Timothy S Chang. « Defining the distance between diseases using SNOMED CT embeddings ». *Journal of Biomedical Informatics* 139 (1 mars 2023): 104307. <https://doi.org/10.1016/j.jbi.2023.104307>.
- (33) European Commission - European Commission. « Un espace européen des données de santé pour les personnes et pour la science ». Text. Consulté le 16 septembre 2023. https://ec.europa.eu/commission/presscorner/detail/fr/ip_22_2711.
- (34) <https://www.health.belgium.be/fr/terminologie-et-systemes-de-codes-snomed-ct>.
- (35) LOI n° 2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé (1), 2002-303 § (2002).
- (36) <https://www.nouvelobs.com/rue89/rue89-temoignage/20130523.RUE6386/a-montpellier-les-degats-de-l-hopital-numerique.html>.
- (37) Scheinker, David, et Margaret L. Brandeau. « Implementing Analytics Projects in a

- Hospital: Successes, Failures, and Opportunities ». *INFORMS Journal on Applied Analytics* 50, n° 3 (mai 2020): 176-89. <https://doi.org/10.1287/inte.2020.1036>.
- (38) Acker, Dominique. « L'organisation des soins primaires dans les pays nordiques ». *Revue française des affaires sociales*, n° 1 (2020): 313-23. <https://doi.org/10.3917/rfas.201.0313>.
- (39) https://medias.vie-publique.fr/data_storage_s3/rapport/pdf/134000136.pdf
- (40) Décret n° 2023-251 du 3 avril 2023 relatif au dossier pharmaceutique, 2023-251 § (2023).
- (41) « Dossier pharmaceutique ». Consulté le 16 septembre 2023. <https://www.service-public.fr/particuliers/vosdroits/F16033>.
- (42) « Mon Espace Santé : une bonne notoriété mais des freins à l'appropriation (bilan France Assos Santé) ». Consulté le 16 septembre 2023. <https://www.healthandtech.eu/fr/article/view/15627/espace-sante-bonne-notoriete-mais-freins-appropriation-bilan-france-assos.html>.
- (43) <https://www.healthandtech.eu/fr/article/view/15578/uk-nhs-scotland-lance-national-digital-platform-equivalent-ecossais-espace.html>.
- (44) Agence du Numérique en Santé. « HDS ». Consulté le 16 septembre 2023. <https://esante.gouv.fr/produits-services/hds>.
- (45) « COFRAC - Comité Français d'Accréditation », 11 septembre 2023. <https://www.cofrac.fr/>.
- (46) « Haute Autorité de Santé - Comprendre la certification pour la qualité des soins ». Consulté le 16 septembre 2023. https://www.has-sante.fr/jcms/c_411173/fr/comprendre-la-certification-pour-la-qualite-des-soins.
- (47) Health Data Hub. « Appel à Projets : accompagnement et soutien à la constitution d'entrepôts de données de santé hospitaliers ». Consulté le 15 septembre 2023. <https://www.health-data-hub.fr/aap-eds>.
- (48) https://esante.gouv.fr/sites/default/files/media_entity/documents/doctrine--technique-du-numerique-en-sante_version-2019-comparee_vf.pdf
- (49) « HIPAA Privacy Rule », 16 mars 2023. <https://www.cdc.gov/nhsn/hipaa/index.html>.
- (50) Proofpoint. « Qu'est-ce que la conformité à l'HIPAA (HIPAA Compliance) ? | Proofpoint FR », 31 mai 2021. <https://www.proofpoint.com/fr/threat-reference/hipaa-compliance>.
- (51) <https://www.congress.gov/107/plaws/publ56/PLAW-107publ56.pdf>
- (52) *Le Monde.fr*. « Le Patriot Act, une législation d'exception au bilan très mitigé ». 12 janvier 2015. https://www.lemonde.fr/les-decodeurs/article/2015/01/12/le-patriot-act-une-legislation-d-exception-au-bilan-tres-mitige_4554570_4355770.html.
- (53) « Medcost : Rapport sur la santé dans le monde 2000 ». Consulté le 16 septembre 2023. https://www.medcost.fr/html/economie_sante_eco/eco_070700.htm.
- (54) https://www.prosperity.com/download_file/view_inline/4789
- (55) « L'augmentation des décès inattendus rappelle le besoin d'anticiper ce risque aux urgences », 23 février 2023. <https://www.hospimedia.fr/actualite/analyses/20230223-gestion-des-risques-l-augmentation-des-deces-inattendus>.
- (56) https://esante.gouv.fr/sites/default/files/media_entity/documents/ans_certsante_rapport_public_observatoire_signalements_issis_2022_vf.pdf
- (57) *Le Monde.fr*. « A Paris, la maternité des Bluets frappée par une attaque informatique ».

- 19 octobre 2022. https://www.lemonde.fr/pixels/article/2022/10/19/a-paris-la-maternite-des-bluets-frappee-par-une-attaque-informatique_6146539_4408996.html.
- (58) Actualités, Publié par Violaine Champetier de Ribes |. « L'Estonie, un État numérique performant | France-Estonie ». Consulté le 16 septembre 2023. <https://www.france-estonie.org/estonie-un-exemple-europeen-detat-numerique-performant/>.
- (59) Agence du Numérique en Santé. « Sesali ». Consulté le 16 septembre 2023. <https://esante.gouv.fr/produits-services/sesali>.
- (60) https://esante.gouv.fr/sites/default/files/media_entity/documents/sesali_guide_utilisateur.pdf
- (61) « Welcome to the HL7 FHIR Foundation ». Consulté le 16 septembre 2023. <https://fhir.org/>.
- (62) « OHDSI – Observational Health Data Sciences and Informatics ». Consulté le 16 septembre 2023. <https://www.ohdsi.org/>.
- (63) Cloudflare. « Open Systems Interconnection | Définition Du Modèle OSI ». Consulté le 16 septembre 2023. <https://www.cloudflare.com/fr-fr/learning/ddos/glossary/open-systems-interconnection-model-osi/>.
- (64) « GA4GH ». Consulté le 16 septembre 2023. <https://www.ga4gh.org/>.
- (65) « Expérimentations “article 51” : le rôle clé des données du SNDS pour les évaluations (analyse Afros) ». Consulté le 16 septembre 2023. (67) <https://www.healthandtech.eu/fr/article/view/16741/experimentations-article-51-role-cle-donnees-snds-evaluations-analyse.html>.
- (66) « Le “benchmark” de l’hébergement du Health Data Hub est communicable si caviardé (Cada) (actualisation) ». Consulté le 16 septembre 2023. <https://www.apmnews.com:443/story.php?objet=398916>.
- (67) « Expérimentations “article 51” : le rôle clé des données du SNDS pour les évaluations (analyse Afros) ». Consulté le 16 septembre 2023. <https://www.healthandtech.eu/fr/article/view/16741/experimentations-article-51-role-cle-donnees-snds-evaluations-analyse.html>.
- (68) A, DICOM_Marie, et DICOM_Marie A. « Enregistrement des résultats de tests antigéniques dans SI-DEP ». Ministère de la Santé et de la Prévention, 16 septembre 2023. <https://sante.gouv.fr/soins-et-maladies/maladies/maladies-infectieuses/coronavirus/professionnels-de-sante/article/enregistrement-des-resultats-de-tests-antigeniques-dans-si-dep>.
- (69) Baker, Monya. « 1,500 Scientists Lift the Lid on Reproducibility ». *Nature* 533, n° 7604 (1 mai 2016): 452-54. <https://doi.org/10.1038/533452a>.
- (70) *Le Monde.fr*. « La biologie française minée par des manquements à l’intégrité scientifique ». 23 octobre 2018. https://www.lemonde.fr/sciences/article/2018/10/23/la-biologie-francaise-minee-par-l-inconduite-scientifique_5373162_1650684.html.
- (71) « arXiv.org e-Print archive ». Consulté le 16 septembre 2023. <https://arxiv.org/>.
- (72) Xu, Zhenxing, Jingyuan Chou, Xi Sheryl Zhang, Yuan Luo, Tamara Isakova, Prakash Adekkanattu, Jessica S. Ancker, et al. « Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks ». *Journal of Biomedical Informatics* 102 (1 février 2020): 103361. <https://doi.org/10.1016/j.jbi.2019.103361>.
- (73) « Vers des bases de données entièrement générées par ordinateur pour des études RWE sans patients ? » Consulté le 16 septembre 2023. <https://www.healthandtech.eu/fr/article/view/16247/vers-bases-donnees-entierement-geneeres-ordinateur-etudes-rwe-sans-patients.html>.
- (74) Mohsen, Farida, Hazrat Ali, Nady El Hajj, et Zubair Shah. « Artificial Intelligence-Based Methods for Fusion of Electronic Health Records and Imaging Data ». *Scientific*

- Reports* 12, n° 1 (26 octobre 2022): 17981. <https://doi.org/10.1038/s41598-022-22514-4>.
- (75) Collins, Dr Francis. « Artificial Intelligence Accurately Predicts Protein Folding ». NIH Director's Blog, 27 juillet 2021. <https://directorsblog.nih.gov/2021/07/27/artificial-intelligence-accurately-predicts-protein-folding/>.
- (76) LOI n° 2021-1017 du 2 août 2021 relative à la bioéthique (1), 2021-1017 § (2021).
- (77) Bourcier, Danièle, et Primavera De Filippi. « Transparence des algorithmes face à l'open data : quel statut pour les données d'apprentissage ? » *Revue française d'administration publique* 167, n° 3 (2018): 525-37. <https://doi.org/10.3917/rfap.167.0525>.
- (78) « Législation sur l'intelligence artificielle: le Conseil appelle à promouvoir une IA sûre et respectueuse des droits fondamentaux ». Consulté le 16 septembre 2023. <https://www.consilium.europa.eu/fr/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/>.
- (79) « Éthique et gouvernance de l'intelligence artificielle pour la santé : orientations de l'OMS : résumé analytique ». Consulté le 16 septembre 2023. <https://www.who.int/fr/publications-detail/9789240037403>.
- (80) Ekins, S, J Mestres, et B Testa. « In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling ». *British Journal of Pharmacology* 152, n° 1 (septembre 2007): 9-20. <https://doi.org/10.1038/sj.bjp.0707305>.
- (81) Nazarova, Antonina L., et Vsevolod Katritch. « It all clicks together: In silico drug discovery becoming mainstream ». *Clinical and Translational Medicine* 12, n° 4 (4 avril 2022): e766. <https://doi.org/10.1002/ctm2.766>.
- (82) Okegbile, Samuel D., Jun Cai, Dusit Niyato, et Changyan Yi. « Human Digital Twin for Personalized Healthcare: Vision, Architecture and Future Directions ». *IEEE Network* 37, n° 2 (mars 2023): 262-69. <https://doi.org/10.1109/MNET.118.2200071>.
- (83) Tulip. « Jumeaux numériques : Un guide pour les fabricants ». Consulté le 16 septembre 2023. <https://tulip.co/fr/blog/digital-twins-in-manufacturing/>.
- (84) <https://www.3ds.com/fileadmin/PRODUCTS-SERVICES/DELMIA/PDF/Whitepaper/DELMIA-APRISO-Digital-Twin-Whitepaper.pdf>
- (86) Haute Autorité de Santé. « Entrepôts de Données de Santé Hospitaliers : La HAS Publie Un Panorama Inédit En France ». Consulté le 16 septembre 2023. https://www.has-sante.fr/jcms/p_3386076/en/entrepots-de-donnees-de-sante-hospitaliers-la-has-publie-un-panorama-inedit-en-france.