HAUTE AUTORITÉ DE SANTÉ

**RAPPORT**

# Entrepôts de données de santé hospitaliers en France

## Quel potentiel pour la Haute Autorité de santé ?

**Validé par le Collège le 20/10/2022**

# Descriptif de la publication

| Titre | Entrepôts de données de santé hospitaliers en France |
|---|---|
| **Méthode de travail** | Le rapport s'appuie sur des éléments de travaux publiés dans la littérature et des entretiens menés avec des experts. |
| **Objectif(s)** | Comprendre ce que sont les entrepôts de données de santé hospitaliers en France : leur répartition dans les CHU, les modes d'organisation, les données qu'ils contiennent et les usages qu'ils servent. |
| **Cibles concernées** | Professionnels, organisations et institutions |
| **Demandeur** | Autosaisine |
| **Promoteur(s)** | Haute Autorité de santé (HAS) |
| **Pilotage du projet** | Mission Data (Pierre-Alain Jachiet) |
| **Recherche documentaire** | Mireille Cecchin (documentaliste) |
| **Auteurs** | Équipe projet de la mission data : Matthieu Doutreligne (chef de projet), Adeline Degremont (cheffe de projet), Pierre-Alain Jachiet (chef de la mission data)<br><br>Groupe d'appui méthodologique : Xavier Tannier, Antoine Lamer |
| **Conflits d'intérêts** | Les membres du groupe de travail ont communiqué leurs déclarations publiques d'intérêts à la HAS. Elles sont consultables sur le site https://dpi.sante.gouv.fr. Elles ont été analysées selon la grille d'analyse du guide des déclarations d'intérêts et de gestion des conflits d'intérêts de la HAS. Les intérêts déclarés par les membres du groupe de travail ont été considérés comme étant compatibles avec leur participation à ce travail. |
| **Validation** | Version du 20 octobre 2022 |
| **Actualisation** | |
| **Autres formats** | |

Ce document ainsi que sa référence bibliographique sont téléchargeables sur www.has-sante.fr

Haute Autorité de santé – Service communication et information
5, avenue du Stade de France – 93218 SAINT-DENIS LA PLAINE CEDEX. Tél. : +33 (0)1 55 93 70 00

# Sommaire

# Résumé

Les données de vie réelle suscitent un intérêt croissant des agences sanitaires à travers le monde, que ce soit pour étudier l'usage, l'efficacité et la sécurité des produits de santé, suivre et améliorer la qualité des soins, réaliser des études épidémiologiques ou faciliter la veille sanitaire. Parmi les différentes sources de données, les entrepôts de données de santé hospitaliers (EDSH) connaissent actuellement un développement rapide sur le territoire français. Dans la perspective de mobiliser ces données dans le cadre de ses missions, la Haute Autorité de santé (HAS) a souhaité mieux comprendre cette dynamique et le potentiel de ces données. Elle a initié en novembre 2021 un travail de recherche visant à dresser un état des lieux des EDSH en France.

Ce rapport pose d'abord le cadre et des définitions en s'appuyant sur la littérature. Il détaille ensuite la méthodologie de recherche, fondée sur des entretiens menés auprès d'acteurs impliqués dans les EDSH de 17 CHU et 5 autres établissements hospitaliers. Le résultat de ces entretiens est structuré par thématiques : historique, gouvernance, données intégrées, usages couverts, transparence, architecture technique et qualité de la donnée. Ce rapport discute ensuite les points d'attention identifiés pour le bon développement des EDSH et des usages secondaires des données. Il ébauche enfin deux cas d'usages pertinents pour la HAS.

La mise en place d'un EDSH constitue un projet complexe, qui implique la collaboration de plusieurs directions de l'hôpital, ainsi que l'appui de cliniciens, d'académiques et d'acteurs industriels. L'équipe opérationnelle entrepôt est pluridisciplinaire, réunissant des compétences pointues dans les domaines médicaux, informatiques et règlementaires.

La transparence vis-à-vis des patients sur les usages de l'EDSH se traduit souvent par des portails publics d'études en cours.

Les données intégrées à l'EDSH sont des extractions automatisées depuis les systèmes d'information hospitaliers. Les difficultés pour réutiliser les données de l'EDSH sont donc directement liées à la complexité de ces systèmes sources. Malgré l'hétérogénéité sur le territoire des solutions logicielles sources, il existe un socle commun des catégories de données intégrées dans les EDSH. Celui-ci comporte les données administratives des patients, les diagnostics et procédures codées, la biologie et les données textuelles. Les données sur le circuit du médicament sont également très souvent intégrées. C'est plus rarement le cas pour des types de données complexes et volumineuses comme la réanimation, l'imagerie ou la génomique.

Les premiers EDSH ont été conçus pour faciliter la construction de l'information hospitalière pour le financement ou le pilotage des établissements et pour des usages proches du soin. C'est aujourd'hui la finalité de recherche qui motive la construction et le développement des EDSH.

L'architecture technique des EDSH contient les couches de traitement, de stockage et d'exposition de la donnée auxquelles s'ajoutent des fonctionnalités annexes comme la gestion des identités, la journalisation ou la gestion des ressources de calculs. Selon le panel d'usages couvert par l'EDSH, et son degré d'ouverture vers l'extérieur de l'hôpital, la plateforme technologique est plus ou moins complexe.

Des processus de mise en qualité des données sont initiés, mais pas systématiquement automatisés via des programmes informatiques et rarement documentés. Malgré la diffusion du modèle OMOP pour les études multicentres, aucun modèle commun de données ne fait actuellement consensus. Les transformations de données depuis les SI sources vers les jeux de données d'étude sont peu documentées publiquement.

# 1. Introduction, motivation

## 1.1. Un intérêt croissant des agences sanitaires pour les données en vie réelle

> **Les données de vie réelle sont les données collectées en pratique courante. Elles peuvent provenir de plusieurs types de sources : dossiers patients informatisés (*Electronic Medical Records*), dossiers de santé informatisés (*Electronic Health Records*), bases médico-administratives (dont la facturation), registres, cohortes, données générées par les patients (ex. : questionnaires, données mobiles ou appareils à domicile)** (1-4)**.**

Les données de vie réelle apportent des informations importantes pour décrire les conditions d'utilisation des produits et technologies de santé, en mesurer la sécurité, l'efficacité ou l'utilité en pratique courante. Elles peuvent également permettre d'apprécier l'impact organisationnel des produits et technologies de santé (2, 5). Si les agences sanitaires utilisent couramment ce type de données pour contextualiser l'évaluation des produits de santé (2, 6), l'exploitation croissante des bases de données et l'accélération des développements cliniques ont mis en avant le potentiel de ces données pour évaluer l'efficacité et la tolérance en vie réelle des produits et technologies de santé. Ces dernières années, les agences sanitaires de nombreux pays ont mené des travaux de fond pour mieux accompagner la génération de données en vie réelle et leur utilisation (1-4). Plusieurs programmes d'études ont été lancés par les agences régulatrices : le programme DARWIN EU par l'Agence européenne du médicament (7) et le *Real World Evidence Program* par la *Food and Drug Administration* (8). La HAS s'est également réorganisée en créant une cellule de coordination sur les données de vie réelle auprès de la direction de l'évaluation et de l'accès à l'innovation pour mieux accompagner les avancées récentes de ce type de données.
Au-delà des questions liées aux produits de santé, les données de vie réelle permettent aussi de mesurer la qualité, sécurité ou pertinence des soins, de mener des études épidémiologiques observationnelles, de faire de la veille sanitaire et du pilotage des soins aux niveaux local et national (9). La crise de la Covid-19 a vu se multiplier ces derniers usages (10, 11).

### 1.1.1. Les entrepôts de données de santé (EDS)

> **Les entrepôts de données de santé (EDS) désignent la mise en commun des données d'un ou plusieurs systèmes d'information médicaux, sous un format homogène pour des réutilisations à des fins de pilotage, de recherche ou dans le cadre des soins.**

Les registres spécialisés ou les enquêtes patients ont un apport important sur des questions de recherches précises, ce qui justifie un effort spécifique de collecte. *A contrario*, les données médico-administratives et les dossiers patients informatisés (DPI) sont collectés en routine. Ces sources couvrent une grande diversité de patients sur des temps longs. Elles sont donc particulièrement intéressantes car elles permettent de répondre à de multiples questions de recherche.

En pratique, la possibilité de mobiliser ces données collectées en routine dépend beaucoup de leur degré de concentration, dans un gradient qui va de la centralisation dans un système d'information (SI) unique et homogène, à l'éclatement dans une multitude de SI aux formats hétérogènes. La structure des SI est un reflet de la

structure de gouvernance. Ainsi, la facilité à travailler sur ces données dépend fortement de l'organisation des acteurs du soin.

Certains pays concentrent les acteurs du soin au sein d'un petit nombre d'organisations. Ce type de structuration aboutit à des sources de données de vie réelle uniformes. Ainsi, en Israël, le plus gros fournisseur de services de santé (Clalit) assure et soigne plus de la moitié de la population. En Corée du Sud, les SI de l'agence gouvernementale responsable de la performance du système de soin et de la qualité (HIRA) sont connectés à ceux de tous les acteurs du soin. L'Angleterre a un système de soin centralisé sous l'égide du *National Health Service*. Cette organisation lui a permis de réunir les données de médecine de ville en deux grandes bases de données qui correspondent aux deux grands éditeurs de logiciels. Actuellement, une première plateforme d'exploitation pour la recherche sur la Covid-19 existe[1] et devrait être suivie par d'autres plateformes similaires pour des thématiques plus généralistes.

À l'inverse, la production de données de vie réelle peut être répartie entre de nombreuses entités, sans gouvernance commune, ayant fait des choix technologiques différents. C'est le cas des systèmes d'assurances et des hôpitaux aux États-Unis. Le regroupement des assureurs en grandes entités permet néanmoins d'aboutir à de larges bases de données comme Medicare[2], Medicaid[3] ou IBM MarketScan[4]. L'Allemagne a fait le constat de systèmes de collecte de la donnée très hétérogènes, limitant le potentiel des données de santé. Au travers du *Medical Informatics Initiative* (12), elle a créé en 2018 quatre consortiums afin de développer les solutions techniques et organisationnelles permettant d'améliorer l'homogénéité des données cliniques.

En France, les données de facturation sont centralisées dans une base unifiée, opérée par l'assurance maladie, qui constitue la base principale du système national de données de santé (SNDS) (9). Ces données sont bien connues et exploitées depuis de nombreuses années, que ce soit par les agences de santé ou par des organismes de recherche (cf. la cartographie de l'écosystème SNDS de la Plateforme des Données de Santé [5]). Malgré leur intérêt central, les données de facturation du SNDS sont insuffisantes pour certains sujets, par manque d'informations cliniques. Ces dernières sont produites par un grand nombre d'acteurs et ne sont pas intégrées dans une base unifiée.

En ville, de rares sources de données cliniques existent. Des entreprises privées suivent des panels de médecins de ville (Iqvia LPD[6], Thin[7]), dont elles centralisent les données dans des bases unifiées. Le Collège national des généralistes enseignants porte actuellement une initiative d'entrepôt de données en médecine générale. À l'hôpital, des efforts sont menés depuis une dizaine d'années dans certains centres, pour constituer des entrepôts de données à partir des dossiers médicaux électroniques (13-22). Ces travaux connaissent une accélération récente, avec un début de structuration à l'échelle régionale et nationale. Des réseaux régionaux de coopération se mettent en place, avec une première autorisation de la Commission nationale de l'informatique et des libertés (CNIL) en février 2022 pour les projets de l'Ouest DataHub[8]. Le ministère de la Santé et de la

[1] https://www.opensafely.org/

[2] https://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/LimitedDataSets/StandardAnalyticalFiles

[3] https://www.cms.gov/Research-Statistics-Data-and-Systems/Computer-Data-and-Systems/MedicaidDataSourcesGenInfo/miniMAX

[4] https://www.ibm.com/products/marketscan-research-databases/databases

[5] https://ecosysteme-snds.health-data-hub.fr/

[6] https://www.iqvia.com/locations/united-kingdom/library/fact-sheets/iqvia-longitudinal-patient-data

[7] https://www.cegedim-health-data.com/cegedim-health-data/thin-the-health-improvement-network/

[8] https://www.chu-hugo.fr/accueil/2022/02/10/la-cnil-delivre-au-ouest-datahub-la-plateforme-de-donnees-de-sante-du-gcs-hugo-lautorisation-de-mener-ses-premiers-projets-de-recherche-academique/

Prévention a ouvert en juillet 2022 un appel à projets doté de 50 millions d'euros pour mettre en place et renforcer un réseau d'entrepôts de données de santé hospitaliers coordonnés avec la Plateforme des Données de Santé (HDH) d'ici 2025[9].

**Ce travail s'intéresse à cette dernière catégorie d'entrepôts de données de santé hospitaliers (EDSH).**

### 1.1.2. Intérêt pour la HAS

La grande quantité d'information clinique présente dans les hôpitaux est aujourd'hui peu exploitée en dehors de l'usage primaire du soin. Dès lors que des efforts sont effectués pour structurer ces données, elles ont un potentiel important pour tous les acteurs du monde de la santé. La HAS peut avoir intérêt à les mobiliser dans le cadre de ses missions d'évaluation des produits et technologies de santé, d'élaboration de recommandations des bonnes pratiques ou de mesure et d'amélioration de la qualité et de la sécurité des soins. De plus, il est probable que la HAS ait de plus en plus régulièrement à évaluer des études mobilisant ce type de sources de données.

Fin 2020, l'institution s'est dotée d'une stratégie data, avec un axe central sur l'usage des données de vie réelle. La HAS mobilise déjà depuis de nombreuses années le SNDS pour contextualiser ses travaux d'évaluation ou pour mesurer la qualité et la sécurité des soins. Il s'agit dans un premier temps de consolider ces usages, notamment pour étudier les pratiques des professionnels en lien avec la production des recommandations. Ayant identifié les entrepôts de données de santé hospitaliers comme une source potentielle afin d'apporter des éléments cliniques, il a été prévu d'explorer leur usage dans un second temps.

Fin 2021, la mission data de la HAS a initié avec les services un travail pour identifier les cas d'usages pertinents à mener sur les données des entrepôts hospitaliers. Elle s'est cependant rapidement confrontée à un manque d'informations sur ces entrepôts. Avec l'aval du Collège de la HAS, la mission data a donc entrepris de dresser un panorama des entrepôts de données de santé hospitaliers en France, en s'intéressant à plusieurs aspects : les typologies de données, les usages développés localement, les outils techniques mobilisés, les modes de gouvernance, les efforts de documentation et de transparence, ainsi que les processus de mise en qualité des données.

## 1.2. Cadre et définitions

### 1.2.1. Le système d'information hospitalier (SIH)

L'informatisation des dossiers patients papier a conduit les hôpitaux français à se doter au cours des années 2000 de solutions de dossiers médicaux électroniques ou dossiers patients informatisés (DPI). Ceux-ci sont définis comme la collection longitudinale de données de santé dans un système d'information (SI) électronique (23). Utilisé en routine par les cliniciens, le DPI leur permet de consigner et d'interroger les informations cliniques nécessaires pour la prise en charge des patients. Un DPI peut être propre à une institution ou bien partagé entre plusieurs acteurs (24). Ce système central est accompagné d'autres applications métiers telles que la gestion administrative des malades (GAM), la prescription informatisée, les logiciels de biologie, de réanima-

[9] https://www.health-data-hub.fr/actualites/aap-entrepots-donnees-de-sante-hospitaliers

tion ou encore d'imagerie. L'ensemble de ces logiciels constitue le système d'information hospitalier (SIH). Selon le degré de maturité du SIH, les différentes sources de données communiquent plus ou moins bien les unes avec les autres.

### 1.2.2. Entrepôt de données de santé hospitaliers (EDSH)

On peut distinguer trois phases dans la structuration des données pour des réutilisations secondaires, distinctes de l'utilisation initiale pour le soin (25).

Les données sont d'abord **collectées** depuis les différentes sources constituant le SIH. Cette première étape technique de copie permet de centraliser ces données initialement cloisonnées dans chacun des SI. Elle permet également d'effectuer des opérations de traitement dans des environnements adaptés, sans risque d'affecter le fonctionnement premier de ces SI pour le soin. Une fois cette phase effectuée, les données ont changé d'environnement SI et sont chargées dans l'EDSH.

La deuxième étape de **transformation** permet d'intégrer, d'harmoniser et de mettre en qualité ces données (13). Les schémas et les concepts des données provenant des différents systèmes sont rarement homogènes. Il y a donc un effort important de transformation et d'agrégation afin d'aboutir à un entrepôt exploitable. Les données de l'entrepôt (*data warehouse* en anglais) désignent en général le jeu de données obtenu après cette étape. Mais le terme d'entrepôt de données est plus large et désigne également la plateforme technologique utilisée pour transformer ces données.

Enfin, il est nécessaire de **mettre à disposition** des jeux de données spécifiques (parfois nommés *datamarts*) à chaque usage secondaire de la donnée. Dans le cadre de la recherche, ceux-ci sont un sous-ensemble du jeu de données principal, ne contenant que la population d'intérêt. Dans le cadre de réutilisations pour le pilotage, l'organisation des soins ou l'amélioration du SIH, ce sont des vues adaptées au nouveau cadre d'utilisation. Des transformations spécifiques à l'usage secondaire peuvent être réalisées avant la mise à disposition.

La Figure 1 illustre ces trois phases de transformation des données :

- Collecte et copie des sources originales
- Transformation : intégration et harmonisation pour la mise en qualité au sein d'une base centrale :
    - Intégration des sources dans une même base de données
    - Déduplication des identifiants : cette étape concerne non seulement les identifiants des patients, mais aussi ceux des services ou des séjours. Des changements dans le système d'information peuvent avoir eu lieu au cours du temps et nécessiter le regroupement de plusieurs identifiants
    - Standardisation : un modèle commun des données harmonise les différentes sources dans un schéma commun, avec éventuellement des nomenclatures communes
    - Pseudonymisation : suppression des éléments directement identifiants dans les données
- Mise à disposition de jeux de données spécifiques (*datamarts*) pour les réutilisations

Ces étapes correspondent au concept d'ingénierie de la donnée *Extract Transform Load* souvent mentionné dans les publications scientifiques consacrées aux EDSH. Les EDSH existants ont en général une architecture plus complexe que ce schéma théorique. Enfin, la régularité de ces différentes étapes dépend des usages. Dans le cas des études, l'extraction de la population à analyser est généralement réalisée une seule fois. Pour des usages opérationnels, les flux depuis les SI sources sont quotidiens, voire horaires.

Dans ce travail, nous nous intéressons à l'objet technologique qu'est l'EDSH, mais aussi à la structure organisationnelle qui l'opère et aux usages qu'il sert.

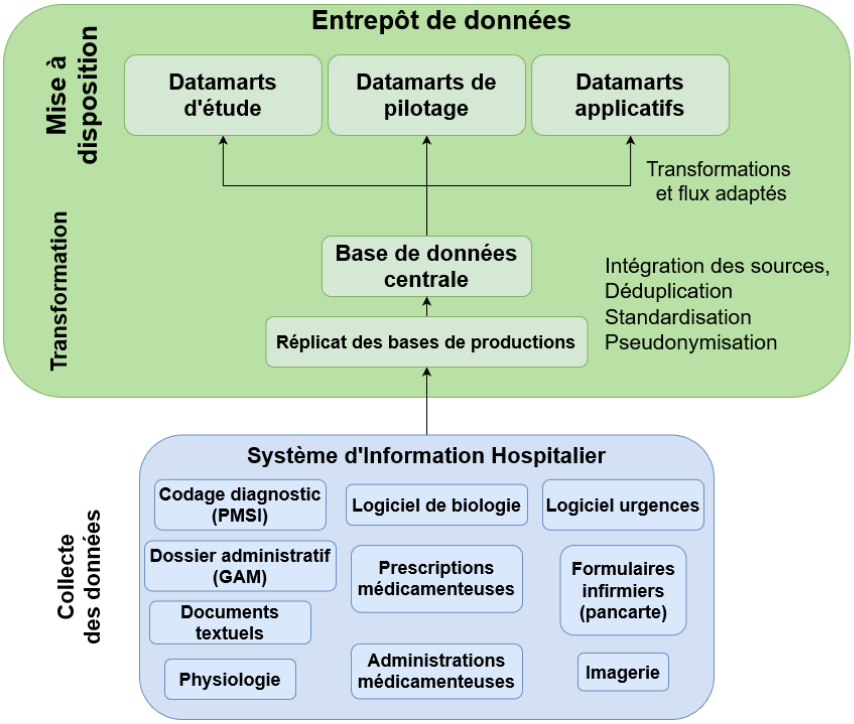


**Figure 1 – Les trois étapes de structuration des données depuis les SI sources : collecte, transformation et mise à disposition**

## 1.2.3. Usage primaire, usage secondaire

Les données de santé sont collectées pour un usage primaire : le soin des patients. Les dossiers médicaux électroniques servent en premier lieu à consigner l'état de santé des patients et leur prise en charge (26). Ces informations peuvent être partagées au sein de l'équipe soignante mais toujours afin de soigner le patient dont les données sont collectées. On dit que l'usage de la donnée d'un patient est primaire lorsqu'elle est utilisée pour la prise en charge de ce même patient[10].

Les usages dits secondaires ne concernent pas directement la prise en charge du patient. Ce sont des réutilisations des données pour d'autres finalités telles que : la recherche, la production d'indicateurs d'activité pour le pilotage ou la qualité des soins, l'optimisation du codage de l'information médicale, la réalisation d'études de faisabilité (26, 27).

Certaines finalités peuvent englober ces deux types d'usages : c'est le cas des systèmes apprenants donnant lieu à des outils d'aide au diagnostic ou de sélection de patients similaires (par leurs traitements et leurs pathologies) (15). Dans ce cas, la donnée d'un ensemble de patients est réutilisée (usage secondaire) pour créer la connaissance nécessaire au fonctionnement de l'outil (phase d'apprentissage). En production, l'outil mobilise les données d'un patient particulier pour donner un résultat dans le cadre de sa prise en charge (usage primaire).

[10] https://ec.europa.eu/commission/presscorner/detail/en/qanda_22_2712

### 1.2.4. Études multicentriques, monocentriques

Si les données sont étudiées par l'équipe soignante des patients au sein d'un service de médecine, on parle d'étude monocentrique. À l'inverse, si plusieurs services, voire plusieurs centres hospitaliers sont impliqués, l'étude est qualifiée de multicentrique.

La multiplication des études multicentriques engendre une complexité accrue sur les plans juridiques (procédures d'habilitation pour des personnels non hospitaliers) et techniques (pseudonymisation, standardisation des données et des nomenclatures, infrastructures de calculs appropriées). Une pseudonymisation est réalisée afin de masquer les informations directement identifiantes aux équipes de recherches. Ces mesures sont nécessaires lorsque les données d'un patient sont manipulées par d'autres acteurs que l'équipe soignante, afin d'en garantir la sécurité, ainsi qu'un usage éthique et scientifique. De manière analogue aux études sur données effectuées hors d'un EDSH, chaque projet de recherche sur EDSH est habilité par un comité scientifique et éthique.

### 1.2.5. Modèle standard de données

Un modèle standard de données, aussi appelé modèle commun, standardise différentes sources dans un même format afin d'assurer la comparabilité des informations collectées par différents organismes (8). Le processus de standardisation implique une correspondance vers un ensemble de tables de données, avec des colonnes communes. Dans sa version la plus aboutie, un modèle standard définit également les nomenclatures utilisées, c'est-à-dire la liste des codes utilisables dans chaque cellule de données, et leur signification. Pour chaque type de données, il est nécessaire de choisir une nomenclature utilisée pour tout l'entrepôt. La granularité de la nomenclature choisie peut entraîner une perte de finesse de l'information présente dans la nomenclature initiale.

L'adoption d'un modèle standard permet d'améliorer la production et le partage de connaissances entre organisations dont les personnels sont exposés aujourd'hui à une très forte mobilité.

Parmi les modèles communs internationaux les plus utilisés en santé, on peut citer I2B2 (28, 29), OHDSI-OMOP (30, 31), Sentinel (32, 33) ou PCORnet (34, 35).

# 2. Méthodes et matériaux collectés

Pour réaliser ce panorama, nous avons sollicité des entretiens auprès des acteurs hospitaliers, qu'ils soient utilisateurs ou parties prenantes des EDSH. Nous nous sommes intéressés à la fois aux EDSH constitués et aux démarches d'EDSH en cours. Nous avons également interrogé des acteurs extérieurs faisant partie de l'écosystème des EDSH : institutions publiques, startups, association. Ces échanges ont eu lieu de mars à juillet 2022.

## 2.1. Recherche des acteurs interrogés

### 2.1.1. Sollicitation de différents acteurs de l'écosystème

Nous avons identifié les acteurs à solliciter selon plusieurs canaux :

- par connaissance de leur rôle institutionnel ;
- en listant sur Légifrance les autorisations CNIL pour un entrepôt de données de santé ;
- par mise en contact à partir des premiers entretiens menés ;
- via l'association d'ingénieurs en données de santé InterHop[11] ;
- grâce à des recherches bibliographiques sur les bases Medline, Embase, Emcare (termes exacts de la recherche en Annexe 1) ;
- grâce au référencement effectué par le ministère de la Santé pour l'appel à projets sur les entrepôts de données de santé[12].

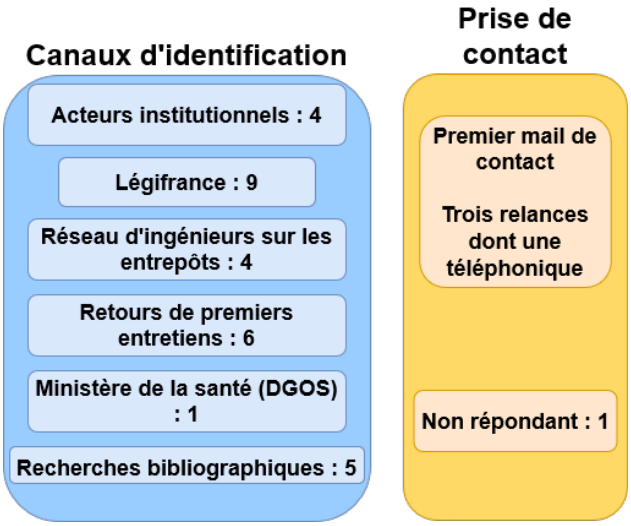


**Figure 2 – Nous avons interrogé 29 organismes repérés via six sources différentes et avons obtenu une réponse de tous sauf un**

Au total, nous avons interrogé 60 participants provenant de 28 organisations : 17 centres hospitalo-universitaires (CHU), 4 centres hospitaliers (dont 2 privés), 3 institutions, 1 centre de lutte contre le cancer (CLCC), 1 association et 2 startups (cf. Acteurs interrogés, liste exhaustive pour la liste complète et les répartitions par

[11] https://interhop.org/
[12] https://solidarites-sante.gouv.fr/archives/archives-presse/archives-communiques-de-presse/ article/ouverture-d-un-appel-a-projets-dote-de-50-millions-d-euros-pour-accompagner-et

pôles de compétence). Pour chaque acteur identifié, nous avons pris contact par courriel, puis réalisé deux relances, ainsi qu'une sollicitation téléphonique en cas de non-réponse.

### 2.1.2. Focalisation sur les gros acteurs hospitaliers

Les résultats portent sur 22 organismes hospitaliers dont 17 CHU. La Figure 3.1 présente les 22 groupes hospitaliers interrogés. Tous opèrent ou construisent un EDS. Parmi eux, quatre sont encore en phase de prototypage.

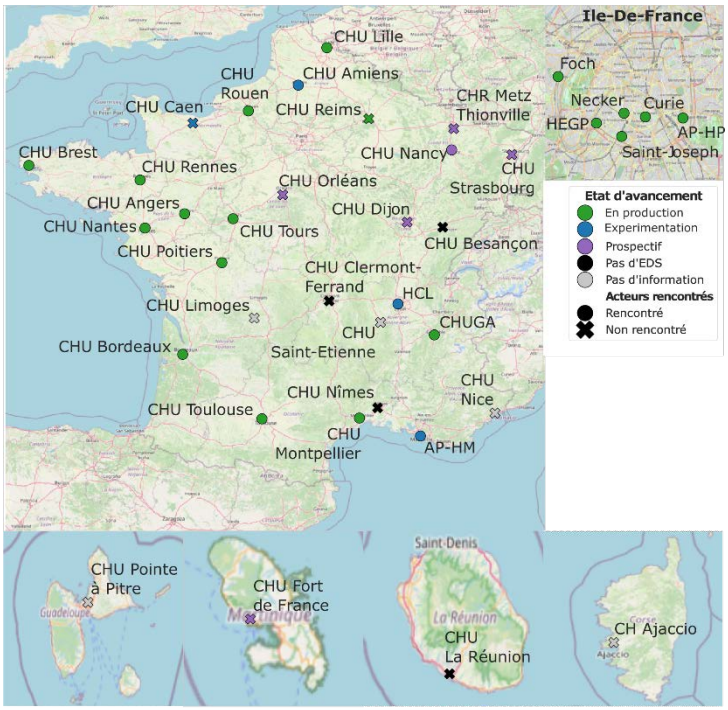


**Figure 3 – Cartographie des organismes interrogés**

## 2.2. Entretiens

Nous avons rencontré les différents acteurs qui participent à la construction et la gestion des EDSH ou qui utilisent cet outil technologique.

Les entretiens ont été menés selon une méthode d'entretien semi-directif à réponses libres. Un formulaire d'entretien a été défini au début du travail, avec des questions abordant les thématiques principales suivantes : l'initiation et la construction de l'EDSH ; l'état des lieux et les projets menés ; les opportunités et obstacles ; les critères de qualité pour la recherche observationnelle ; la faisabilité de deux sujets d'intérêt pour la HAS ; un échange libre. Le formulaire complet, avec les questions précises, est disponible en annexe Acteurs interrogés, liste exhaustive.

Le formulaire d'entretien a été envoyé aux participants en avance pour cadrage, puis a servi de support pour mener les entretiens. Les entretiens ont duré 90 minutes. Ils ont été enregistrés afin de s'y référer lors de la synthèse et de la rédaction de ce rapport.

## 2.3. Méthode d'analyse

### 2.3.1. Aspects quantitatifs

Nous avons choisi de présenter trois tableaux de résultats, dont les champs sont restés stables tout au long du cycle d'entretien (cf. Annexe 6 pour les champs précis). Les deux premiers tableaux portent sur les caractéristiques des acteurs interrogés et sur ceux des entrepôts de données. Nous les avons complétés à partir des notes prises lors des entretiens, des enregistrements, et parfois en relançant les intervenants pour obtenir des compléments d'information.

Le troisième tableau porte sur les études en cours dans les EDSH. Nous avons collecté la liste de ces études sur les portails de déclaration dédiés, que nous avons trouvés pour 10 EDSH sur 18 opérationnels. Puis nous avons élaboré une classification des études, en nous basant sur la typologie des études rétrospectives décrite par le consortium de recherche OHDSI (30). Nous avons enrichi et précisé cette typologie, en la confrontant à la liste des sujets d'études collectés. La nomenclature retenue comporte quatre catégories générales et six catégories fines.

- **Épidémiologie descriptive** que l'on sépare en **chiffrage** ou **caractérisation de population**. Ces études s'intéressent à une population cible bien définie médicalement qu'il s'agit de décrire statistiquement. On parle d'étude de chiffrage lorsque l'on s'intéresse à une variable simple, comme l'incidence ou la prévalence, et d'étude de caractérisation lorsqu'il s'agit d'aller plus loin en étudiant un ensemble de covariables. Les études de faisabilité, ou de pré-screening pour l'inclusion dans une étude, entrent notamment dans ces catégories (36).

- **Épidémiologie analytique** où l'on a séparé la **recherche de facteurs de risque** de **l'évaluation d'un effet de traitement**. Ces deux types d'études tentent de dégager des mécanismes systématiques propres à une pathologie d'intérêt. Elles ont donc un but plus universel que la description d'une situation donnée. Les études des **facteurs de risque** s'intéressent à une cible clinique bien définie (évolution d'une maladie, évènement de soin). Puis, elles cherchent les covariables qui sont le plus associées à cette cible. C'est une étude d'association sans quantification de l'effet causal des facteurs sur la variable d'intérêt. À l'inverse, les études évaluant **l'effet d'un traitement**, correspondant à une intervention bien définie sur une variable précise, cherchent à montrer un lien causal entre ces deux variables (37).

- Mise au point de **processus d'aide à la décision.** Le but de ces études est l'amélioration ou l'automatisation d'un processus diagnostique ou pronostique, à partir des données cliniques d'un patient donné. En général, la finalité est la construction d'un score de risque ou de prévention, voire la mise en place d'un système d'aide au diagnostic. Ces études s'inscrivent dans la démarche d'une médecine individualisée, avec comme horizon un retour au dossier d'un seul patient.

- **Informatique médicale :** certaines études ont un caractère méthodologique ou concernent le développement d'un outil. Elles visent à améliorer la compréhension et la capacité d'action des chercheurs et des cliniciens. Ce type d'étude comprend l'évaluation d'un outil d'aide à la décision, l'extraction d'information depuis des données non structurées, la recherche de méthode de phénotypages automatiques.

Les études ont été classées selon cette nomenclature à partir de leur titre et de leur description.

Nous avons également classé chaque étude selon la spécialité de l'investigateur principal. Nous avons pour cela utilisé la liste des spécialités médicales règlementaires, à laquelle nous avons ajouté certaines directions métiers pouvant être à l'initiative des études : la DRCI, la DSI et l'équipe EDSH.

### 2.3.2. Aspects qualitatifs

Nous avons synthétisé les résultats obtenus selon les grands thèmes du document d'entretien : gouvernance, transparence, données, usages, architecture technique et qualité des données. Des sous-thématiques ont été constituées progressivement par apport d'éléments nouveaux de la part des acteurs interrogés.

# 3. Résultats

## 3.1. Gouvernance et acteurs

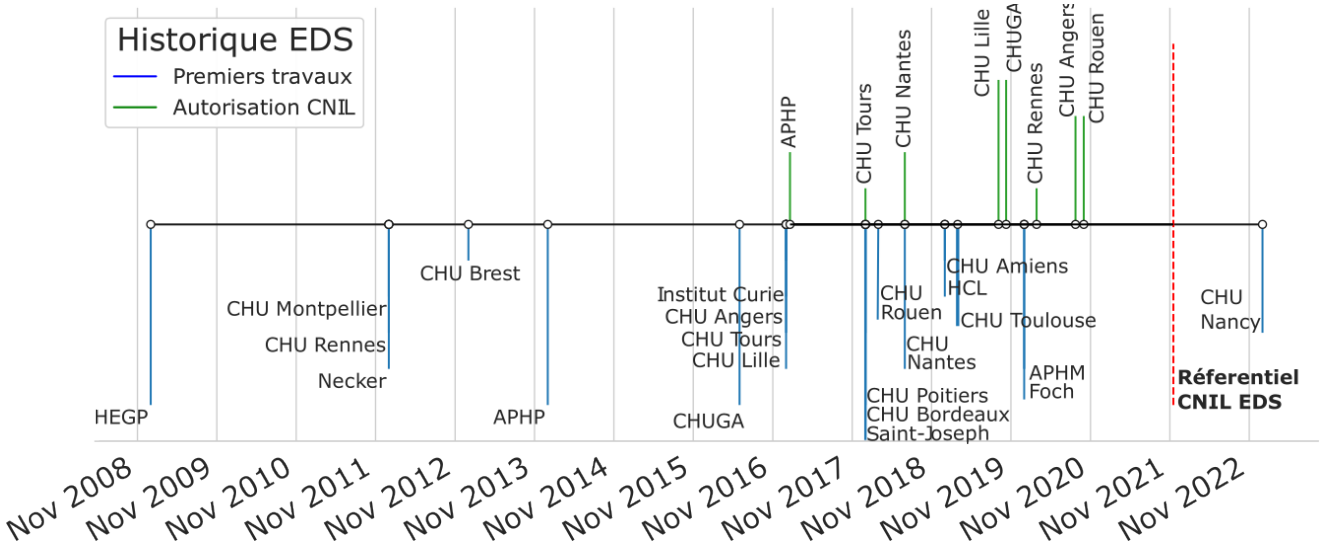


**Figure 4 – La mise en place des entrepôts de données de santé en France remonte à la fin des années 2000 et s'accélère à la fin des années 2010. Les autorisations règlementaires par la CNIL (traits verts) succèdent à de premiers travaux techniques plus anciens (traits bleus)**

### 3.1.1. Principaux acteurs

La Figure 4 permet d'apprécier l'historique de la mise en place des EDSH. Il faut distinguer les premiers travaux (en bleu) qui précèdent systématiquement l'autorisation règlementaire de la CNIL (en vert). En traits pointillés rouges, nous avons indiqué la date de publication du référentiel CNIL ayant fixé le cadre juridique des EDS (27).

Les EDSH ont jusqu'ici été initiés par une ou deux personnes provenant du monde hospitalier. Ces personnes ont une formation universitaire en bio-informatique, informatique médicale ou statistiques. L'initiative de la construction vient très souvent de l'intérieur de l'organisation avec un *cavalier seul* qui convainc progressivement les autres parties prenantes de l'intérêt de la réutilisation des données. La pérennisation de l'EDSH est accompagnée par la construction d'un environnement de coopération entre les différents acteurs (département d'information médicale, direction des systèmes d'information, direction de la recherche clinique et de l'innovation, usagers cliniciens) et un soutien de la direction ou de la commission médicale d'établissement. Elle est accompagnée également par la constitution d'une équipe, ou entité, dédiée au maintien et à l'opérationnalisation de l'EDS. Les initiatives plus récentes comme celles des HCL ou du Grand Est se distinguent par un portage initial plus institutionnel et haut niveau.

L'EDSH a un potentiel fédérateur pour les différentes directions métiers de l'hôpital. La participation active de la DRCI, la DSI et du DIM en fait un espace de collaboration important. Les liens de l'équipe EDSH avec la DRCI sont souvent bien marqués car les EDSH sont mobilisés pour la recherche. Le portage par la DSI permet d'inscrire l'EDSH dans le paysage informatique de l'hôpital. Ce portage est hétérogène selon la couverture fonctionnelle plus ou moins large de l'EDSH et le profil des porteurs initiaux. Le DIM est souvent impliqué et permet d'apporter ses connaissances pointues sur le circuit historique de la donnée. S'il existe systématiquement une équipe EDSH opérationnelle, les ressources humaines qui lui sont attribuées sont en revanche extrêmement variables : d'une moitié d'équivalent temps plein jusqu'à 70 personnes pour l'AP-HP, avec une médiane à 7,5

équivalents temps plein. L'équipe constituée comprend systématiquement un médecin coordinateur. Elle est multidisciplinaire avec des compétences en santé publique, informatique médicale, informatique (service web, base de données, réseau, infrastructure), ingénierie de données et statistiques. Malgré la présence de l'équipe EDSH, il peut être difficile pour un acteur extérieur d'identifier l'interlocuteur principal pour les sujets data, notamment dans les structures de taille intermédiaire (selon les deux startups interrogées).

Des connaissances et techniques à l'état de l'art dans la gestion de multiples formats de données sont utiles pour les EDS. Ces connaissances sont dans le monde universitaire, notamment dans les domaines de la santé publique, des biostatistiques, du traitement du signal, du traitement automatique des langues ou du traitement de l'image. Certains EDSH arrivent à mobiliser ces compétences via des collaborations étroites, tandis que d'autres non. Cette variabilité s'explique plus largement par l'insertion plus ou moins grande des groupes hospitaliers dans les tissus académiques locaux. L'apport de ces compétences permet de développer plus rapidement des usages différents sur l'EDS, exploitant le potentiel de chaque type de données.

### 3.1.2. Interactions avec les cliniciens

Des cliniciens travaillant avec les données de l'EDSH apportent leur aide pour améliorer la qualité des données et mieux qualifier les usages. Une large part des intervenants a souligné la nécessité d'avoir une *proximité* avec ceux qui saisissent la donnée dans le service d'information. Cette proximité peut être renforcée grâce à des personnes spécialisées dans un type de données au sein des EDSH : biologie, imagerie, EEG, anatomopathologie, traitement de texte. Une stratégie complémentaire est d'identifier des interlocuteurs privilégiés, familiarisés avec les données et les outils de l'EDSH (sélecteurs de cohorte, documentation des flux et des données). Ces derniers permettent de former plus largement les équipes cliniques moins proches des problématiques et des outils data. Quelques cliniciens sont formés aux outils de base de données ou au langage de programmation statistique (SQL, R, Python), mais ces compétences restent rares parmi les médecins.

### 3.1.3. Gestion des projets de recherche et des demandes

Avant de démarrer, les projets passent par un guichet de demande unique pour le dépôt et le suivi. Le projet est systématiquement analysé par un comité scientifique et éthique. Un outil de suivi est souvent mentionné (14/22), mais son périmètre fonctionnel est difficile à cerner. Il peut aller de la simple autorisation du projet à la mise à disposition automatique des données. Les processus pour démarrer un nouveau projet sur l'EDSH sont rarement documentés publiquement (7/22). Mais ils sont presque toujours communiqués en interne une fois que l'EDSH est fonctionnel.

### 3.1.4. Coopération public/privé

Historiquement, les premiers EDSH se sont appuyés sur des développements de solutions en interne. Plus récemment, des acteurs privés proposent leurs services concernant la mise en place et l'opérationnalisation des EDSH. Aujourd'hui, il n'est pas rare d'observer une coopération entre l'équipe EDSH et un ou plusieurs acteurs privés (12/22). Ces prestations vont d'une expertise technique afin de constituer les flux de données et les nettoyer jusqu'à la livraison d'une plateforme intégrant les différentes étapes de traitement de la donnée.

Certains acteurs mettent en doute la gestion de l'EDSH par délégation des compétences techniques à des acteurs privés. Si l'expertise des outils est trop peu présente au sein de l'EDSH, l'outil ne saurait pas accompagner l'évolution des besoins en données des hôpitaux. Selon eux, il existe un parallèle avec la mise en place des DPI. Certains établissements ont bénéficié d'une offre précoce peu fonctionnelle. À l'inverse, des internalisations de la gestion des DPI ont permis une meilleure maîtrise du SIH et l'alignement avec les usages internes. Les mises en place tardives de solutions commerciales de DPI se sont appuyées sur des solutions technologiques mieux consolidées et l'expertise d'acteurs internes.

Les schémas des données des SI sources sont nécessaires pour la mise en place des flux de données, depuis le SIH vers l'EDSH. La question de la propriété privée de ces schémas de données, notamment dans le cas de gestionnaires de DPI commerciaux, est une difficulté majeure – mentionnée par cinq organisations – pour la mise en place et le développement des EDSH. À défaut d'en disposer, un premier travail de rétro-ingénierie est nécessaire pour inférer ces schémas de données, travail qui doit être maintenu lors des évolutions des SI. Un flou juridique autour de cette propriété privée des schémas complique leur partage entre centres hospitaliers, et plus généralement le partage des codes informatiques et de la documentation des flux et transformation de données.

Mise à part la constitution de l'EDSH, les partenaires privés désirant utiliser les données sont majoritairement des entreprises désireuses d'améliorer le recrutement pour la recherche clinique à partir de l'EDSH, c'est-à-dire d'améliorer les études de faisabilité et le pré-screening.

La mise en place de l'EDSH est portée par un spécialiste des données de santé soutenu par la direction médicale ou administrative. Le maintien opérationnel et la pérennisation passent par la création d'une équipe pluridisciplinaire d'experts et le développement d'un environnement de collaboration. L'EDSH tisse des liens étroits avec d'autres directions métiers : la DSI pour s'intégrer dans le paysage informatique, la DRCI pour s'inscrire dans les activités de recherche de l'établissement, le DIM pour sa connaissance de l'information hospitalière, les pôles cliniques pour définir et porter les cas d'usage.

## 3.2. Transparence

Les études en cours dans les EDSH sont inégalement référencées publiquement sur les sites des hôpitaux. Certains établissements ont des portails d'études complets, quand d'autres répertorient à peine une dizaine d'études sur leur site public tout en mentionnant plusieurs centaines de projets en cours lors des entretiens. Au total, nous avons retrouvé 10 de ces portails sur 18 EDSH en production. Les usages autres que les études scientifiques en cours ne sont que très rarement documentés publiquement.

La publication de la liste des études en cours est très hétérogène selon les établissements, et morcelée entre plusieurs sources : clinicaltrials.gov, le répertoire public de la Plateforme des Données de Santé ou le site internet de l'hôpital hébergeant l'entrepôt.

## 3.3. Données

### 3.3.1. Dépendance au parc de logiciels du SIH

Les données de l'EDSH sont le reflet des SI utilisés au quotidien par le personnel de l'hôpital. Les acteurs soulignent que la qualité des données de l'EDSH et la quantité de travail à fournir pour une réutilisation rapide et efficace sont fortement dépendantes des SI sources. La possibilité d'accéder aux données d'un SI dans un format structuré et normalisé simplifie grandement son intégration dans l'EDS puis sa réutilisation.

Le parc de logiciels d'un établissement et l'organisation locale des soins influencent donc le contenu de l'EDSH.

Pour exemple, l'AP-HP a généralisé le choix du DPI Orbis. Son déploiement s'est étalé de 2004 à 2019, ce qui a abouti à des versions différentes du logiciel sur ses nombreux sites et services. Les données n'étant pas exposées de façon normalisée entre versions, il en résulte une grande complexité des bases de données et des difficultés d'intégration dans l'EDSH. Une des solutions trouvées pour un traitement unifié repose sur l'extraction d'informations à partir de PDF.

Autre exemple, les HCL ont internalisé le développement de leur DPI depuis 2012, qu'ils distribuent dans la région Rhône-Alpes via le groupement d'intérêt HOPSIS. La maîtrise des données du DPI facilite la mise en œuvre d'un EDSH aux HCL, et l'uniformité au niveau régional facilitera la création d'un réseau de collaboration.

### 3.3.2. Catégories de données intégrées

Si le paysage logiciel est varié sur le territoire, les grandes fonctionnalités des SIH sont toutefois les mêmes. On peut donc mener une analyse du contenu des EDSH, selon les grandes catégories de données communes.

Le socle commun indispensable à tous les EDSH est constitué par les données du logiciel de gestion administrative des malades (identification des patients, mouvements hospitaliers) et de la facturation (PMSI). Puis des flux sont progressivement développés à partir des différents logiciels constituant le SIH. Le but est de construire un schéma de données homogène, liant les sources entre elles, maîtrisé par l'équipe EDSH. Toutes les informations du DPI ne sont pas forcément intégrées. La priorisation des sources se fait grâce à des projets thématiques, qui nourrissent la démarche de construction de l'EDS. Ces projets permettent d'améliorer la compréhension des sources impliquées, en confrontant l'équipe EDSH aux problèmes de qualité présents dans les données.

La biologie structurée et les textes sont quasiment tout le temps intégrés (20/22 et 21/22). Les textes contiennent une grande quantité d'informations. Ils constituent des données non structurées donc sont plus difficilement exploitables que des tables structurées. De plus, les types de textes sont nombreux : comptes-rendus de chaque spécialité, ordonnances, lettres de sortie, notes infirmiers, lettres de consultation, réunions de concertation pluridisciplinaire, etc. L'AP-HP intègre ainsi 70 types de textes différents.

Puis, sont intégrés le circuit hospitalier du médicament (prescriptions et administrations, 19/23) et, plus occasionnellement, la réanimation (3/22) ou la pancarte (constantes relevées par les infirmiers, 4/22). L'imagerie est rarement intégrée (5/22), notamment pour des questions de volumétrie. Les données génomiques sont bien repérées, mais jamais intégrées, même si elles sont parfois jugées importantes et inscrites au programme de travail de l'EDSH.

Le Tableau 1 résume les sources de données intégrées dans les EDSH.

| Type de données | Nombre d'EDS | Ratio |
|---|---|---|
| Gestion administrative des malades | 22 | 100 % |
| PMSI | 22 | 100 % |
| Textes | 21 | 95 % |
| Biologie | 20 | 91 % |
| Circuit du médicament | 19 | 86 % |
| Imagerie | 5 | 23 % |
| Pancarte | 4 | 18 % |
| Anatomopathologie | 3 | 14 % |
| Réanimation | 3 | 14 % |
| Dispositifs médicaux | 2 | 9 % |

**Tableau 1 – Sources de données intégrées dans les EDSH**

Les différents EDSH ont intégré un socle commun de données, notamment autour du PMSI, des données administratives, de la biologie, du circuit du médicament et des textes. Les niveaux de qualification sont hétérogènes et difficiles à objectiver sans documentation publique, ni standard commun. La complexité des flux à mettre en place dépend beaucoup de la qualité des logiciels d'information sources. Les données volumineuses (génomiques et imagerie) ne sont à la portée que des gros établissements avec une influence régionale du fait des moyens techniques à mettre en œuvre pour les intégrer.

## 3.4. Usages

### 3.4.1. Triple usage : recherche, pilotage, clinique

Les premiers EDSH ont été à l'origine conçus principalement pour optimiser le codage hospitalier à des fins de facturation, et pour décloisonner les systèmes d'information hospitaliers à destination des soignants. Certains travaillent encore des usages proches du soin : au travers du pilotage (13/22) ou d'une amélioration du dossier patient informatique pour faire de la fouille transverse de données (7/22). En revanche, aujourd'hui, l'usage principalement mis en avant pour la constitution d'EDSH est celui de la recherche scientifique.

Lorsqu'un usage hors recherche est envisagé sur la plateforme de l'EDSH, il est nécessaire de distinguer la plateforme des usages qui y sont menés lors de l'homologation CNIL. En effet, le cadre du référentiel CNIL ne s'applique qu'au cadre strict de la recherche. Au moins un CHU a clarifié les cadres juridiques et techniques adaptés à chaque usage : recherche (pseudonymisation à la source) ou soin (pseudonymisation réversible pour des cas de prévention par exemple).

### 3.4.2. Recherche

L'EDSH est souvent utilisé afin de simplifier la recherche au sein de l'hôpital. Les questions posées sont observationnelles (non interventionnelles). La Figure 5 présente la distribution en 6 catégories (cf. méthode) des 248 études collectées sur les portails d'études en cours de 9 CHU et 1 CH. Les études portent d'abord sur la caractérisation de population (25 %), suivie du développement de processus d'aide à la décision (24 %), l'étude de facteurs de risque (18 %) et l'étude des effets de traitement (16 %).

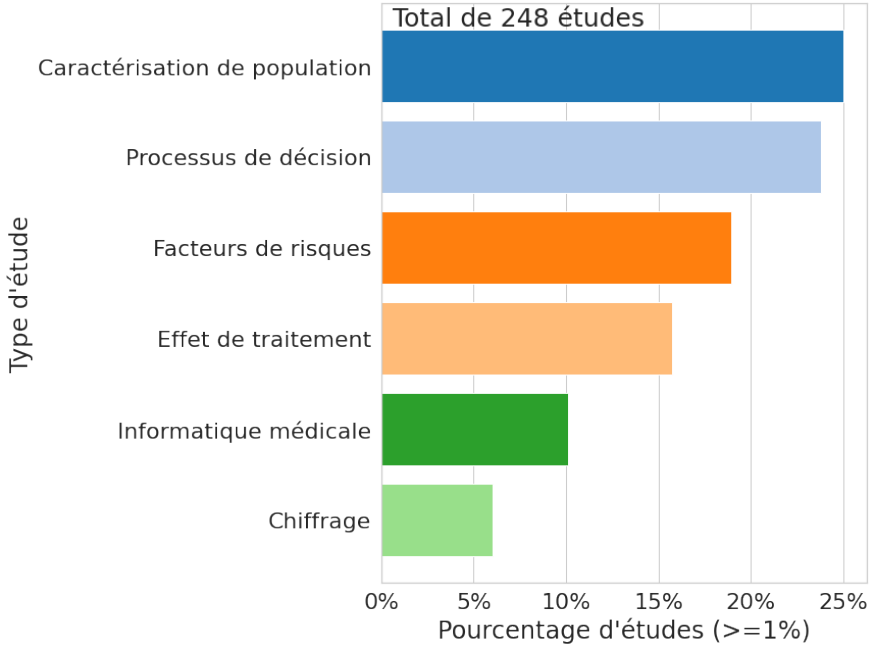


**Figure 5 – Répartition des études en cours dans 9 CHU et 1 CH, par catégorie**

La Figure 6 décrit la répartition par spécialité de l'investigateur principal de l'étude. Quatre spécialités sont surreprésentées : la santé publique, la réanimation, la médecine interne et les pathologies cardio-vasculaires. Les autres spécialités sont moins représentées (répartition exhaustive en Annexe 5).

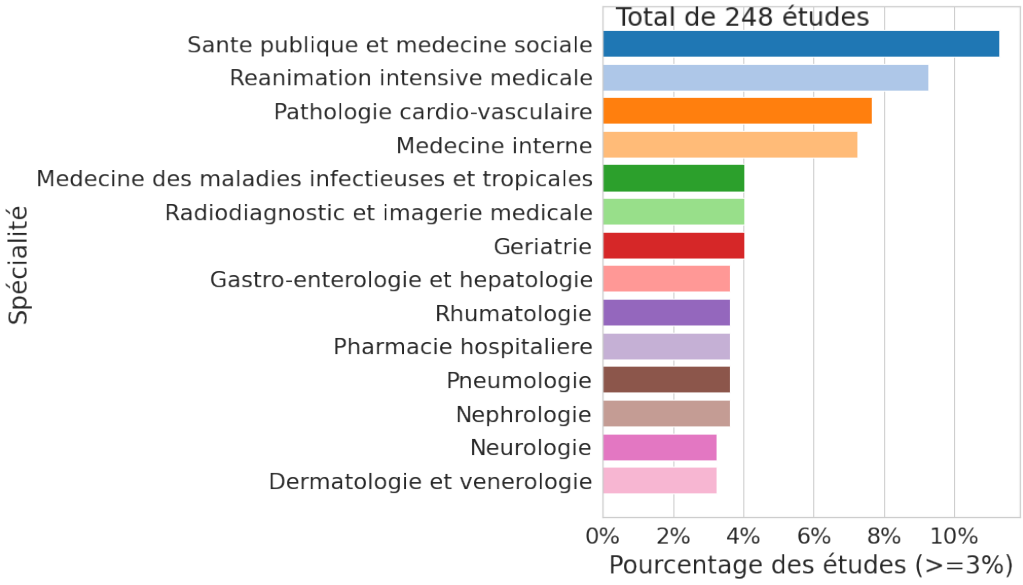


**Figure 6 – Distribution des études par spécialité de l'investigateur principal. Seules les spécialités représentant plus de 3 % des usages sont représentées**

Les EDSH constitués sont beaucoup utilisés pour des projets internes comme les thèses d'internat (au moins dans 9/22) et servent d'infrastructure pour la recherche monoservice : leur grand intérêt étant le désilotage des différents systèmes d'information. Pour la plupart des établissements interrogés, il manque encore des moyens et une maturité des méthodes et outils pour mener des recherches interétablissements (Ouest Data-Hub[13]) ou via les appels à projets européens (EHDEN[14]). Ces deux réseaux sont rendus possibles par une gouvernance supralocale et un schéma de données commun, respectivement eHop (38) et OMOP (31). L'AP-HP, grâce à sa dimension régionale et le choix du standard OMOP, prend également la direction d'une recherche inter-CHU. Parallèlement, la région Grand Est construit un réseau d'EDSH sur le modèle du Grand Ouest en se basant également sur eHop. Enfin, l'uniformité des DPI en Auvergne-Rhône-Alpes favorisa la création d'un réseau inter-EDSH régional.

### 3.4.3. Pilotage

Les EDSH ont parfois été initiés pour faciliter la construction de l'information médicale, afin d'améliorer et d'optimiser le codage PMSI (4/22). Les textes cliniques rassemblés dans une même base de données sont requêtés via des mots clés pour faciliter la structuration d'informations. Les données sont ensuite agrégées dans des indicateurs dont certains remontent au niveau national (PMSI ou financements expérimentaux de l'article 51). Certains acteurs ont développé une grande expertise dans l'extraction d'informations en s'appuyant sur des ontologies ou des systèmes de règles.

La construction d'indicateurs à partir des données cliniques peut également servir à la gestion administrative de l'établissement. Ce pilotage administratif est traditionnellement réalisé indépendamment des données cliniques, ou par des outils différents selon les services, sans homogénéisation des informations remontées au niveau de l'hôpital.

Enfin, concernant davantage la clinique, certains acteurs déclarent que l'EDSH a un potentiel encore trop peu exploité pour le pilotage de l'activité médicale par les services cliniques. Selon eux, l'EDSH pourrait également permettre de faire des retours réguliers et adaptés aux professionnels de santé sur leurs pratiques. Ces retours contribuent à augmenter l'implication et l'intérêt des professionnels de santé dans les projets de l'EDSH.

L'EDSH intéresse parfois pour la veille sanitaire ou la pharmacovigilance (7/22). De nombreux projets de surveillance sanitaire ont notamment été développés lors des premières vagues de Covid-19.

### 3.4.4. Soins, usage clinique

Les EDSH peuvent également servir dans le cadre du soin, grâce à des applicatifs spécifiques qui apportent des fonctionnalités nouvelles par rapport aux logiciels de soins. Des moteurs de recherche permettent d'interroger toutes les données de l'hôpital rassemblées dans l'EDSH, sans cloisonnement entre SI. Des interfaces dédiées peuvent alors offrir une vue unifiée de l'historique des données d'un patient, avec une transversalité interspécialités, précieuse notamment en médecine interne. Ces outils de recherche transverse permettent également aux professionnels de santé de faire de la recherche rapide dans l'ensemble des textes, par exemple pour trouver des patients similaires (15). Des usages de prévention, d'automatisation de tâches répétitives et de coordination des soins sont également mis en avant (11/22). Des exemples concrets sont le tri automatique des prescriptions hospitalières par ordre de complexité, ou la mise en place de filières spécialisées pour la prévention primaire ou secondaire.

[13] https://www.chu-hugo.fr/accueil/projets/Ouest-DataHub/

[14] European Health Data Evidence Network – ehden.eu

Les usages de pilotage et de décloisonnement du SIH à destination des soignants existent mais ne sont plus majoritaires. La vague d'intérêt récente est motivée par des objectifs de recherche en épidémiologie, santé publique et aide à la décision médicale. Ce renouvellement est soutenu par une volonté nationale d'étendre l'accès aux données et les usages, notamment pour des projets multicentres ou portés par des acteurs extrahospitaliers : évaluation des produits de santé, espace européen de données de santé, innovation.

## 3.5. Architecture technique

L'architecture technique d'un EDSH comporte plusieurs couches :

- **le traitement de la donnée :** connexion et export des donnes sources, transformations diverses (nettoyage, agrégation, filtrage, standardisation) ;
- **le stockage de la donnée :** moteurs de base de données, stockage des fichiers (sur des serveurs de fichiers ou des *stockages objets)*, moteurs d'indexation permettant d'optimiser certaines requêtes ;
- **l'exposition de la donnée :** données brutes, API, tableaux de bord, environnements de développement et d'analyse, applications web spécifiques.

À ces composants centraux s'ajoutent d'autres briques transverses qui assurent le fonctionnement efficace et sécurisé de la plateforme : gestion des identités et des autorisations, journalisation de l'activité (logging), administration automatisée des serveurs et des applications.

Selon l'EDSH, le nombre de briques techniques utilisées peut varier fortement : 36 logiciels pour la plateforme AP-HP, une dizaine référencée par l'institut Curie[15]. L'intégration de ces différentes technologies constitue la plateforme de stockage et d'exposition des données de l'EDSH. Des éditeurs proposent de telles plateformes intégrées. La partie amont de traitement de la donnée est assez spécifique à chaque EDS, bien que des outils se développent, portés par certains prestataires.

Il n'y a pas d'homogénéité des briques technologiques utilisées bien que certains acteurs partagent des solutions similaires – soit commerciales, soit *open source* via le réseau d'ingénieurs InterHop[16] ou des dépôts de codes hospitaliers[17].

Les briques *open source* partagées sont des dépôts de code destinés à certains usages précis : traitement des dates, fichiers de correspondances de nomenclatures, solution de questionnaires électroniques, déploiement de datalab, interface de recherche et de sélection de patients. Ces outils permettent de s'appuyer sur l'expérience des autres EDSH, d'économiser du temps et de mutualiser des compétences pointues. Ils nécessitent des compétences internes en informatique pour être adaptés au contexte local et interfacés les uns avec les autres. La démarche d'ouverture de ces outils est portée par les développeurs. Plusieurs acteurs ont mentionné des difficultés pour justifier l'investissement nécessaire pour rendre ces outils plus génériques et mieux documentés, de façon à faciliter leur réappropriation par d'autres EDSH.

L'environnement de développement et d'analyse (datalabs Jupyterhub ou RStudio) est un élément clé de la plateforme, car il permet de traiter les données au sein de l'infrastructure de l'EDSH. Tous n'en disposent pas à la date de notre étude (7/22) ; cependant, presque tous ont décidé de s'en doter en vision cible. Actuellement, les équipes de recherche clinique travaillent encore souvent sur des extractions des données, dans des environnements moins sécurisés.

[15] https://artifacthub.io/packages/search?page=1&repo=curie-df-helm-charts

[16] https://framagit.org/groups/interhop/

[17] https://github.com/aphp/

L'architecture des EDSH est difficilement auditable car les solutions techniques sont hétérogènes et la documentation rarement disponible. De plus, selon le nombre d'usages que cherche à couvrir l'entrepôt, la complexité de l'infrastructure technique peut varier. Les centres ayant fait le pari des études multicentriques ont des architectures plus sophistiquées couvrant un plus large panel d'usages.

## 3.6. Qualité de la donnée, formats et méthodes standards

### 3.6.1. Suivi de la qualité de la donnée

Des processus de suivi systématique de la qualité des données sont en cours de construction dans certains EDSH. Souvent (11/22), des scripts tournent à intervalle régulier afin de détecter des anomalies techniques dans les flux de données. Quelques outils d'investigation de la qualité des données, sous forme de tableaux de bord, commencent à être développés en interne (6/22). Des réflexions théoriques sont en cours sur la possibilité d'automatiser des vérifications sur la cohérence des données, par exemple démographique ou temporelle. Certains établissements tirent aléatoirement des dossiers dans le DPI et les comparent avec les informations présentes dans l'EDSH afin d'en mesurer l'exhaustivité.

### 3.6.2. Modèle de données

Aucun modèle de données standard ne se dégage comme utilisé par tous les EDSH. Tous connaissent l'existence des modèles OMOP (standard de recherche) et FHIR (standard de communication).

Plusieurs CHU considèrent le modèle OMOP comme une partie centrale de l'entrepôt, notamment pour les usages de recherche (8/22). Cet engouement a été favorisé par l'appel à projets européen EHDEN [18], lancé par le consortium de recherche OHDSI à l'origine de ce modèle de données.

Dans l'ouest de la France, les EDSH utilisent le logiciel d'entrepôt eHop. Ce dernier utilise un modèle commun de données nommé également eHop. Cela permet au consortium d'EDSH appartenant au groupement de coopération sanitaire HUGO de lancer des projets régionaux ambitieux. Ce modèle est amené à s'étendre avec le futur réseau d'entrepôts du Grand Est ayant également choisi cette solution. En comptant ce regroupement et les autres établissements ayant fait le choix d'eHop, ce modèle compte 12 établissements sur les 32 CHU. Néanmoins, eHop ne définit pas de nomenclature standard à utiliser dans son modèle. Des travaux importants sont en cours pour aboutir à un modèle sémantique commun via des nomenclatures de références.

### 3.6.3. Construction de variables

Dans le cadre de la recherche clinique traditionnelle, l'analyse est conduite à l'aide d'une table patient afin de répondre à une question de recherche précise. Chaque ligne correspond à un patient et chaque colonne de la table contient une information pertinente sur celui-ci dans le cadre de l'étude.

Au contraire, les données contenues dans un entrepôt sont extrêmement variées. À un patient donné correspondent généralement des centaines de données : gestion administrative, consultations, prélèvements de biologie, relevés physiologiques, prescriptions médicamenteuses, etc. Pour chaque question d'étude, il est donc

[18] https://www.ehden.eu/

nécessaire d'extraire un sous-ensemble d'informations pour aboutir à une table patient pertinente. Cela nécessite de structurer et de normaliser des variables, avec une difficulté plus ou moins grande selon que l'on parte de données structurées, comme la biologie, ou d'informations textuelles contenues dans les notes cliniques.

Un autre enjeu est de pouvoir reproduire ces analyses d'un centre à l'autre afin de comparer des résultats issus d'une même méthode de travail. Ceci est particulièrement important pour dégager des résultats avec une forte validité externe. Il est donc nécessaire de développer des méthodes d'extraction de variables qui soient robustes et réplicables.

Dans la majorité des EDSH, l'extraction des variables est le résultat d'un dialogue entre utilisateurs et collecteurs de la donnée. Cette démarche n'est quasiment jamais systématisée en une méthodologie susceptible d'être répliquée d'une étude à l'autre. Cependant, de premières méthodologies sont en cours d'élaboration afin d'établir des méthodes de référence (39).

### 3.6.4. Documentation

La moitié des EDSH ont mis en place une documentation accessible au sein du CHU sur les flux de données, la signification ainsi que le bon usage des données qualifiées (12/22 mentionnés). Deux établissements nous ont parlé de catalogues de données, qui documentent l'ensemble des sources présentes dans l'EDSH et leur provenance. Cette documentation sert à l'équipe qui développe et maintient en conditions opérationnelles l'entrepôt. Elle sert également aux utilisateurs pour comprendre les transformations effectuées sur les données. Cette documentation n'est en revanche jamais ouverte à l'extérieur de l'entrepôt. Aucun schéma des données une fois celles-ci transformées et préparées pour l'analyse n'est publié. Quelques EDSH publient des schémas généraux sur les grandes étapes de transformation de données, depuis les sources brutes vers les tables d'exploitation, avec les principales briques technologiques impliquées (5/22).

Le contrôle qualité des données est majoritairement réalisé grâce à la mise en place d'études et le retour des cliniciens sur des données manquantes ou aberrantes. Des tableaux de bord hebdomadaires sont parfois réalisés pour contrôler les flux de données. L'adoption de modèles de données et de nomenclatures communes est inégale. Globalement, les EDSH publient très peu d'informations à destination des acteurs externes, notamment concernant l'architecture, les flux ou les bonnes pratiques d'utilisation des données de l'entrepôt à des fins de recherche, d'innovation ou de régulation. En revanche, une documentation interne semble très souvent exister.

# 4. Discussion

## 4.1. Points d'attention et opportunités

### 4.1.1. Gouvernance

#### Constituer une équipe dédiée, pluridisciplinaire et transverse

Dès lors que l'EDSH devient une composante essentielle de la gestion des données à l'hôpital, il faudrait encourager la constitution d'une équipe autonome consacrée à l'architecture de données, l'automatisation des processus et la documentation des données (40). Cette équipe devrait développer une excellente connaissance du processus de collecte des données et des réutilisations potentielles afin de qualifier les différents flux provenant des SI sources, les standardiser vers un schéma homogène et harmoniser la sémantique. Son rôle devrait également inclure la diffusion des connaissances produites sur les données hospitalières auprès des chercheurs usagers.

Cette équipe devrait privilégier la pluridisciplinarité, avec des experts dans chaque domaine : médical, informatique, science des données, statistique. Pour assurer la transversalité, il faudrait idéalement que des personnels soient détachés depuis le DIM, la DSI et la DRCI au sein de l'équipe entrepôt. À l'inverse, des personnels référents au sein de ces directions peuvent jouer le rôle d'interlocuteurs privilégiés, en se formant spécifiquement aux outils et processus de l'entrepôt. Afin d'établir un esprit de confiance, il apparaît important d'informer chaque direction lorsque ses données sont réutilisées pour des projets particuliers. Il ne s'agit pas d'une demande d'autorisation, car les données n'appartiennent pas à une direction ou un service particulier.

#### L'internalisation pérenne de compétences est nécessaire mais difficile

Les EDSH ne sont pas des outils techniques figés ; ils devraient évoluer avec le SIH et les usages. Une grande partie des acteurs ont souligné la nécessité d'internaliser les compétences permettant de maintenir et faire évoluer les solutions initiales (11/22). Les acteurs ont partagé leur crainte d'une externalisation des compétences en sciences des données, qui réduirait leur capacité d'action.

Tous les acteurs interrogés déclarent des difficultés de ressources humaines. Les EDSH mobilisent des compétences techniques de haut niveau, et l'expérience acquise dans chaque contexte devrait être capitalisée sur le long terme. Il est à la fois difficile d'attirer des profils expérimentés et de garder des profils juniors plus de quelques années, notamment pour des questions de valorisation professionnelle ou financière. Il est par ailleurs difficile d'intégrer des cliniciens capables d'apporter leur expertise sur l'intégration des données brutes depuis le SIH. Les soignants manquent de temps disponible à consacrer à ces sujets et n'ont pas la même culture scientifique de la donnée que les porteurs des EDSH. Il est essentiel de diriger une partie des usages vers ces soignants (entre autres pour le pilotage de leurs activités) afin de les intéresser au projet d'EDSH.

Les ressources propres à l'entrepôt sont rares et souvent prises sur d'autres budgets, ou sur des crédits par projet. Si ce fonctionnement est naturel pour une phase initiale de prototypage, il ne semble pas adapté au caractère pérenne et transversal de l'outil. L'étendue des usages couverts laisse penser que l'EDSH devient progressivement une infrastructure importante pour la recherche et le pilotage des soins. Les acteurs demandent donc plus de visibilité sur le long terme afin de pouvoir construire des équipes et des projets pérennes. Cette problématique a été identifiée par le comité stratégique des données de santé, qui a mis en place un groupe de travail dédié au financement des bases de données de santé, avec un premier travail spécialement consacré aux EDSH.

### Des modes de gouvernance qui convergent mais une lisibilité institutionnelle à améliorer

Le référentiel CNIL requiert la création d'un comité de pilotage et d'un comité scientifique et éthique. Ce texte a également précisé la composition et les rôles de ces comités. Cette clarification a permis une homogénéisation progressive des modes de gouvernance sur le territoire (27). Néanmoins, l'entrepôt étant souvent rattaché à plusieurs directions métiers, l'interlocuteur privilégié vers qui se tourner est parfois difficile à identifier pour un acteur extérieur. Un espace de site internet dédié à l'EDSH dans chaque centre améliorerait la lisibilité de l'organisation. Le référencement de ces espaces dans un catalogue national des EDSH permettrait un suivi de ce paysage en évolution rapide.

### Mettre en place une gouvernance à trois niveaux

La gouvernance de l'EDSH se joue idéalement à trois niveaux : local au sein du CHU, interrégional et national.

Le niveau local semble indispensable pour comprendre la collecte des données et porter des usages pertinents du point de vue des soignants.

Le niveau interrégional semble adapté à des collaborations multicentres pérennes. Les Groupements interrégionaux pour la recherche clinique et l'innovation (GIRCI) sont à la base de la gouvernance territoriale de la recherche appliquée en santé[19]. C'est un niveau intéressant pour mutualiser certaines compétences rares et pointues nécessaires à chaque EDSH. La spécialisation locale peut être orientée vers un type de données (biologie, réanimation, médicaments) ou une brique technique spécifique de l'EDSH. L'inscription dans un réseau de coopération permet à chaque acteur de bénéficier de ces compétences et des investissements spécifiques effectués localement. L'interrégional apparaît également le niveau auquel il est possible de se coordonner sur des solutions techniques homogènes et interopérables, en exploitant notamment les similarités des SIH. Cette échelle ou un niveau plus fin (par exemple régional) semble pertinent pour la collecte de données dans une réflexion territoriale de coordination des soins. Ces données peuvent être hébergées dans un entrepôt interrégional ou national. Une alternative à la mise en commun des données dans une plateforme tierce est un système de requêtes fédérées. Dans ce cas, les résultats des requêtes sont agrégés à partir des données présentes dans chaque EDSH local. Que ce soit un entrepôt central ou une architecture fédérée, il est nécessaire auparavant de standardiser les schémas et les nomenclatures de données.

Enfin, le niveau national semble être l'échelon adapté pour coordonner les acteurs, mettre à disposition des ressources et des outils communs à tous. Ceux-ci peuvent être d'ordre technique : infrastructures sécurisées, plateforme technique, brique technologique spécifique (par exemple, outils pour la pseudonymisation). Sur le plan méthodologique, c'est également ce niveau qui semble adapté pour spécifier les schémas de données communs et les nomenclatures employés pour les usages nationaux ou européens. Le niveau national permet également d'impulser et de mener des coopérations sur des sujets thématiques pointus. C'est aussi à cet échelon qu'un appui juridique semble particulièrement intéressant. Celui-ci pourrait consister à mieux définir le cadre législatif des EDSH et les types de contractualisations entre gestionnaires de SI et EDSH.

[19] https://solidarites-sante.gouv.fr/systeme-de-sante-et-medico-social/innovation-et-recherche/l-innovation-et-la-recherche-clinique/article/groupements-interregionaux-pour-la-recherche-clinique-et-l-innovation-girci#contacts

**Un exemple de réseau régional d'EDS, l'Ouest DataHub**

Dans le Grand Ouest, les responsabilités du projet de la plateforme Ouest DataHub sont réparties entre les différents CHU, créant un espace propice à l'échange et à la coopération. Nantes est responsable de la plateforme technologique. Rennes contribue avec son expertise historique sur les études observationnelles. Tours est spécialiste sur les dispositifs médicaux. Angers apporte un expert en sécurité informatique. Brest a imaginé le modèle d'organisation locale – le centre de données clinique. Ce partage des tâches ne limite pas l'autonomie de chaque EDSH. Chaque centre peut proposer un projet interrégional et en être l'investigateur principal.

#### Toutes les structures hospitalières n'ont pas la capacité de se doter d'un EDSH

L'importance de l'usage recherche et les investissements nécessaires à la création d'un EDSH sont particulièrement adaptés aux CHU et aux autres grands centres hospitaliers. En revanche, ces mêmes facteurs rendent peu crédible leur développement dans chaque structure hospitalière. Une éventuelle exhaustivité nationale de la couverture des EDSH sur le champ hospitalier ne serait atteinte que par une homogénéisation des SIH et de leurs structures de données.

### 4.1.2. Transparence

#### Malgré des efforts constants, l'information scientifique et patient est encore insuffisante

Les recommandations internationales (1, 10, 25) sont en faveur d'un référencement public des projets en cours, avec une publication préalable des protocoles de recherche, essentielle du point de vue scientifique pour contrôler les biais. Tous les établissements publient une partie de leurs études sur clinicaltrials.gov dans la catégorie *recherches observationnelles*. Ce registre ne distingue pas les études menées sur entrepôts. Étant donné le changement d'échelle permis par les EDSH (en termes d'accès à des données variées ou de puissance statistique), il est regrettable que la distinction ne soit pas explicite entre les recherches effectuées avec des données extraites à la main dans un unique service et celles qui utilisent les données de tout un centre hospitalier, voire d'une région entière.

Il existe également deux sortes de portails d'information à destination des patients : le registre des projets sur données de la Plateforme des Données de Santé et les portails d'études en cours sur EDSH dans chaque hôpital. Il est dommage d'avoir un doublon de ces systèmes de déclaration des études à destination des patients. Une seule source pourrait être renseignée par les chercheurs et la seconde alimentée automatiquement à partir de la source de référence, en s'accordant sur des métadonnées communes.

Du point de vue d'un patient, il n'existe à ce jour aucun moyen de savoir si ses données personnelles sont incluses pour un projet spécifique. La solution technique semble complexe à mettre en œuvre. Cependant, une information des patients de meilleure qualité sur la réutilisation des données les concernant est nécessaire pour construire une confiance sur le long terme. Pour cela, il serait nécessaire d'établir et de tenir à jour les portails d'études en cours propres à chaque établissement qu'exige le référentiel CNIL.

### 4.1.3. Données

#### Vers la création d'un socle commun de données ?

Un intérêt certain existe dans les différents centres pour mener des études multicentres. Ces études permettent de disposer de la puissance statistique nécessaire pour les pathologies rares. Elles sont également intéressantes afin de généraliser ou contraster les conclusions d'un centre à l'autre.

Dans la majorité des EDSH, il existe déjà un socle commun de données sur lequel mener ce type d'études. Celui-ci concerne les données patients les plus essentielles et parfois celles les plus facilement intégrables : données administratives, PMSI, actes CCAM, physiologie, prescriptions de médicaments, biologie. Cependant, malgré une structuration locale homogène au sein des EDSH, ce patrimoine de données n'est à l'heure actuelle pas mobilisable sans schémas ni nomenclatures communes.

#### Encourager les modèles standards ouverts, éprouvés et internationaux

Afin de disposer d'un socle pertinent pour des usages multicentres, il serait nécessaire de définir des schémas de données partagés et des nomenclatures communes. Il manque actuellement une coordination nationale sur ces sujets, afin d'appuyer les efforts locaux. En effet, que ce soit via l'initiative européenne EHDEN ou pour des projets multicentres, il existe divers efforts de mise en correspondance entre des nomenclatures locales et des nomenclatures de références. Ce travail est long et nécessite des compétences médicales pointues. Le partage des résultats à l'ensemble des acteurs via des dictionnaires et des outils partagés permettrait d'accélérer le processus (via des outils comme Susana[20], développé par InterHop). La mise en commun améliorerait également la qualité des correspondances créées en renforçant le consensus entre les acteurs.

Enfin, il existe des standards ouverts, éprouvés et internationaux. Par exemple, OHDSI spécifie un ensemble de nomenclatures standards, largement utilisé dans le cadre de son réseau de recherche[21]. Il faudrait les utiliser au maximum, en limitant la création de nouveaux modèles de données ou de nouvelles nomenclatures « standards » sur des périmètres déjà couverts.

### 4.1.4. Usages

#### Les recherches multicentres changent le rapport à la donnée

La personne qui analyse la donnée n'a pas défini le processus de collecte des données et ne connaît généralement pas le contexte de saisie des informations dans le SIH. Cette nouvelle dimension de recherche nécessite un développement beaucoup plus grand des compétences en science des données afin de changer les points d'attention auparavant focalisés sur la mise en œuvre du plan d'étude statistique. La réutilisation des données exige de consacrer plus d'efforts à la préparation des données et à la documentation des transformations effectuées.

Les CHU ayant le plus de projets fonctionnels se nourrissent des projets de recherche en cours pour alimenter une démarche d'amélioration continue de l'entrepôt. La confrontation entre les expertises métiers et les équipes entrepôts permet de progressivement intégrer de nouvelles sources, de corriger les anomalies et de mettre en qualité les données pour des usages bien qualifiés. L'idée que les jeux de données peuvent être parfaitement nettoyés une fois pour toutes, pour tous les usages doit être déconstruite. Cette démarche de co-construction des projets de recherche et de l'entrepôt de données a été mentionnée par de nombreux acteurs, notamment par les premiers utilisateurs des EDSH en France.

[20] https://framagit.org/interhop/omop/susana

[21] https://athena.ohdsi.org/search-terms/terms?standardConcept=Standard&page=1&pageSize=15&query=

### Concilier les données de ville et de l'hôpital : un enjeu majeur mais encore lointain

Les questions d'études liées aux trajectoires de soins nécessitent de combiner les données hospitalières et les données de ville, typiquement par un appariement de données de l'EDSH à la base principale du SNDS. Selon deux acteurs, il est difficile de respecter les contraintes techno-juridiques liées au référentiel SNDS (41) lorsque celui-ci est apparié avec les EDSH. Les spécifications de sécurité du référentiel SNDS entraînent notamment une augmentation importante des coûts (traçage des utilisateurs, isolation des réseaux, chiffrements). Afin de respecter ces contraintes, une solution consiste à conduire les travaux dédiés au SNDS au sein d'une seconde infrastructure dédiée. Ceci aboutit à une complexification des flux, des procédures d'accès, à une baisse de l'offre de service, ainsi qu'à un cloisonnement des usages ville/hôpital. En vue d'une convergence de la vision ville/hôpital, il serait judicieux d'harmoniser les référentiels, en prenant en compte les différents types de cas d'usage et niveau de sécurité requis.

Que ce soit pour les usages de recherche, de pilotage ou de soin, la vision transverse ville/hôpital des données de santé est importante. Pour certains acteurs, les cadres de collaboration pour des travaux sur la territorialisation des soins sont encore trop peu développés (avec l'ARS, Santé publique France ou les observatoires régionaux de santé).

### Des exigences juridiques complexes qui se clarifient à l'usage

Le triple usage, recherche, pilotage et soin, crée des problématiques juridiques lors des demandes d'autorisation CNIL concernant la séparation des systèmes d'information et le principe de proportionnalité. L'exemple suivant sur la pseudonymisation illustre la construction progressive du cadre légal des EDS.

#### La pseudonymisation, illustration de la construction progressive du cadre légal des EDS

L'exemple de la pseudonymisation illustre bien la construction progressive du cadre légal et des exigences techniques imposés aux hôpitaux. Pour toute étude effectuée à partir de l'EDSH de l'AP-HP, la CNIL a exigé dès 2017 de pseudonymiser les données mises à disposition des chercheurs (27). Fallait-il pseudonymiser lors de l'ingestion des flux et ainsi ne plus pouvoir remonter à l'identité des patients ? Cela permettait de faire de la recherche multicentres. Mais cela empêchait les autres catégories d'usages nécessitant un retour à l'identité des patients : applications de soins et de recrutement, aide au diagnostic, requêteurs internes à destination des cliniciens, outils de prévention. Le référentiel de novembre 2021 semble avoir clarifié le mode opératoire en laissant la possibilité de réidentification pour certains cas d'usages où c'est indispensable, comme la suppression des données d'un individu à sa demande, l'inclusion dans un essai ou les urgences médicales (27), SEC-REI-1 à 5. Les EDSH s'interrogent actuellement sur la qualité minimale à atteindre pour le système de pseudonymisation. C'est une problématique complexe et pour laquelle aucun consensus n'a actuellement été trouvé en France (42).

### Les études menées sur EDSH bénéficieraient d'une catégorisation par type d'objectif

La classification des types d'études que nous avons établie dans la section 2.3 est une proposition de notre part pour mieux comprendre les objectifs et les méthodes des recherches effectuées à partir de l'entrepôt. En tant qu'observateur extérieur, il nous a été délicat de rendre intelligibles des catégories d'études à partir des simples portails d'études menées sur EDSH. Plus spécifiquement, il était compliqué de faire la part des recherches mobilisant l'entrepôt relevant de la recherche interventionnelle *versus* observationnelle, de distinguer les études cherchant à mieux comprendre une pathologie de celles développant de nouveaux outils pronostiques ou diagnostiques. Il nous paraît important de mettre en place une telle catégorisation, uniforme sur le territoire et renseignée dans les registres d'études, afin de mieux suivre le type de travaux menés sur les EDSH. Une telle catégorisation permettrait d'établir un vocabulaire partagé pour définir les objectifs des études. Elle aurait également pour intérêt de faciliter les échanges entre les acteurs lors des projets, surtout lorsqu'ils viennent de disciplines scientifiques différentes.

### Les réutilisations orientées vers des usages primaires de la donnée sont encore rares

L'usage des données pour améliorer la coordination des soins, leur qualité, ou faire de la prévention est souvent évoqué (11/22). En réalité, ces usages, moins portés par le DIM ou la DRCI, sont encore minoritaires et ne sont pas soutenus par des financements adaptés. En revanche, de nombreuses études en cours concernent des processus d'aide à la décision dont le but est un gain de temps pour les professionnels de santé. Ce sont souvent des projets de recherche, pas encore des utilisations intégrées dans le cadre des soins courants. Quelques indicateurs de qualité sont également étudiés, mais sont rarement restitués localement dans une démarche d'amélioration continue et d'organisation des soins.

## 4.1.5. Architecture technique

### Éparpillement des solutions techniques

Si la gouvernance des EDSH tend à devenir homogène, il n'en est pas de même pour les outils, les méthodes et les formats de données du fait de l'innovation technique forte et de la présence de nombreux acteurs. Comme suggéré par le récent rapport sur l'utilisation des données pour la recherche au Royaume-Uni (40), il serait judicieux de privilégier un petit nombre de plateformes techniques modèles afin de limiter l'hétérogénéité des solutions. Cela est valable pour les briques logicielles (ex. : eCRF, solutions de plateforme data) et les protocoles d'interopérabilité (HL7 FHIR, OMOP CDM).

### Favoriser l'émergence de plateformes technologiques *open source*

La mise en place des EDSH a nécessité des investissements importants pour développer certains outils essentiels à leur fonctionnement, en particulier les plateformes de stockage et d'exposition des données. Certaines de ces plateformes sont devenues des produits commerciaux, par la création d'une entreprise *spin-off* (Codoc pour Dr Warehouse, initialement *open source*) ou l'association avec un industriel (eHop avec Enovacom, filiale santé d'Orange Business Service). Ces plateformes ont ainsi pu être vendues à d'autres hôpitaux, accélérant la mise en place de leurs entrepôts et assurant une forme de mutualisation des investissements dans une solution commune.

Ces premières plateformes commerciales sont distribuées avec des licences propriétaires et un code source fermé, ce qui pose plusieurs difficultés. Une solution fermée manque par nature de transparence : il n'est pas possible d'étudier son fonctionnement pour le comprendre ou le corriger. Elle crée une dépendance forte à l'éditeur, qui doit intervenir pour toute adaptation aux besoins spécifiques de l'hôpital. L'éditeur est maître de la feuille de route des développements, qui peuvent diverger des besoins locaux ; un problème soulevé par certains hôpitaux concernant leurs DPI commerciaux. Les solutions propriétaires proposent plus naturellement des suites cohérentes de logiciels, bien intégrés entre eux. Mais ces logiciels sont moins modulaires et moins interopérables avec des logiciels tiers. Ainsi, une solution propriétaire a tendance à enfermer[22] dans un écosystème d'outils, avec des difficultés de portabilité vers une autre solution si la qualité se dégrade ou si le coût augmente. Ce risque de verrouillage est particulièrement important pour la plateforme de données, qui est un nœud central en interaction avec de nombreux flux en amont, et des outils et des usages en aval.

*A contrario*, les logiciels libres ou *open source* offrent une transparence par défaut. Ils permettent une indépendance des éditeurs et des opérateurs, favorisant une vraie mise en concurrence et prévenant les phénomènes de rente. Lorsqu'une gouvernance adaptée est mise en place, ils respectent une feuille de route et des principes élaborés collectivement. Ils permettent de mutualiser les contributions de plusieurs acteurs, engendrant des économies d'échelle importantes et favorisant la recherche de consensus. La transparence sur les failles de sécurité assure également des mises à jour plus fréquentes que des solutions propriétaires. Ce mode de développement permet de produire des logiciels fiables, de qualité, faciles à interopérer avec

[22] https://fr.wikipedia.org/wiki/Enfermement_propriétaire/

d'autres. Il est intéressant de souligner que l'écrasante majorité des briques logicielles mobilisées pour construire les plateformes données des EDSH, que ce soit pour le stockage et le traitement de données, sont des logiciels *open source*. Pourtant, si plusieurs EDSH insistent sur l'importance de la maîtrise complète de leur plateforme et continuent à investir des ressources localement, il faut noter qu'aucune plateforme centrale n'est aujourd'hui activement développée en *open source*.

Il y a donc une opportunité à favoriser l'émergence de plateformes technologiques *open source* et des communautés associées. La France a déjà un cadre favorable pour l'*open source* avec de fortes retombées économiques[23] (43). Depuis 2012, une circulaire du Premier ministre recommande aux administrations de préférer les logiciels libres dans leur politique d'achat (44). L'impact de cette politique sur le développement de l'*open source* en France a été évalué comme important, avec des effets positifs sur la dynamique économique générale du secteur informatique (45). En 2016, l'article 16 de la loi pour une République numérique encourage à nouveau les administrations publiques à utiliser des logiciels libres et des formats ouverts (46). Elle prévoit par ailleurs une ouverture par défaut pour tous les codes sources écrits dans le cadre d'une mission de service public. Une circulaire du Premier ministre a renouvelé ces ambitions en 2021 (47). De nombreuses ressources interministérielles ont été développées pour accompagner ce mouvement : un pôle d'expertise logiciels libres rattaché à Etalab au sein de la direction interministérielle du numérique (DINUM), un plan d'action logiciels libres et communs numériques, un socle interministériel de logiciels libres[24], etc. La vision est partagée au niveau européen, avec la création de Joinup[25] par la commission fin 2011, pour soutenir les solutions interopérables, ouvertes et libres. Localement, plusieurs EDSH insistent sur l'importance de la maîtrise complète de leur plateforme et continuent à investir des ressources. Les ingénieurs des EDSH sont généralement enthousiastes pour l'*open source* et cherchent à s'organiser en réseau. Certains outils sont partagés en *open source*, ouvrant des possibilités de collaboration, mais sans investissement suffisant – ni gouvernance commune – pour les transposer aisément d'un EDSH à l'autre.

Un premier frein à lever est le manque de soutien institutionnel. Au niveau local, il faudrait favoriser le logiciel libre dans les politiques d'achat. Il faudrait valoriser davantage l'ouverture et la participation à des communautés *open source*. Les bénéfices directs en seront une amélioration de la qualité logicielle et une attractivité renforcée pour le recrutement. Cette forme de *soft power* a été bien comprise par les géants informatiques qui ont des politiques *open source* fortes et des unités internes pour les porter (les *Open Source Program Offices*). Au niveau national, il faudrait développer une politique industrielle qui ferait de l'*open source* un avantage compétitif, quand il semble aujourd'hui plus simple de lever des fonds avec la promesse de clients captifs. L'appel à projets pour les EDSH pourrait ainsi augmenter le financement éligible des coûts liés à la dissémination des connaissances (1 % dans la limite de 50 k€ par bénéficiaire), à un niveau comparable aux dépenses liées à la propriété intellectuelle (30 %).

Il faut également continuer le travail d'acculturation, pour dépasser certaines habitudes et lever certaines craintes. En particulier, le logiciel libre est encore trop souvent associé à la gratuité et à l'absence de garantie. Pourtant, l'écrasante majorité des serveurs informatiques fonctionnent avec le système d'exploitation libre GNU/Linux, tandis que de nombreuses entreprises fondées sur des logiciels *open source* sont valorisées plusieurs (dizaines de) milliards : Wordpress (moteur de 40 % des sites internet), RedHat (éditeur de solutions *open source* d'entreprises), GitLab (forge logicielle), Databricks (plateforme de traitement de données), Hugging Face (*machine learning*), pour n'en citer que quelques-unes.

[23] https://www.bercynumerique.finances.gouv.fr/les-logiciels-libres-et-open-source-en-europe-un-etat-des-lieux, https://labo.societenumerique.gouv.fr/2022/01/27/dossier-les-logiciels-libres-et-open-source-en-france-ou-en-sommes-nous/

[24] https://www.etalab.gouv.fr/accompagnement-logiciels-libres/, https://communs.numerique.gouv.fr/plan-action-logiciels-libres-et-communs-numeriques/, https://sill.etalab.gouv.fr/software

[25] https://joinup.ec.europa.eu/collection/joinup/about

### Des compétences internes pour opérer et faire évoluer les outils au plus proche des besoins

Par ailleurs, il faut démystifier le rôle dominant des outils et insister à nouveau sur l'importance des équipes internes. De bons choix technologiques sont indispensables pour construire des EDSH fonctionnels. En revanche, ces technologies doivent être accompagnées par des experts internes capables de les mettre en œuvre, de les adapter aux spécificités et aux besoins locaux, et de les faire évoluer dans la durée.

Tous les acteurs précisent qu'un intérêt de l'EDSH est d'améliorer la compréhension et l'interprétation des flux de données hospitaliers. Or, cette compréhension est le fruit d'une démarche continue d'échanges entre les cliniciens collectant la donnée et les utilisateurs (chercheurs, administratifs ou soignants), médiés par l'équipe EDSH. L'installation d'une plateforme ne garantit pas de développer cette connaissance des données. Il est donc illusoire d'attendre qu'un acteur extérieur ou interne à l'hôpital construise l'EDSH parfait, réplicable d'un site à l'autre, sans interactions continues entre une équipe EDSH et l'écosystème de l'EDSH.

## 4.1.6. Qualité de la donnée

### Encore trop peu d'équipes qualité de la donnée

La notion de qualité des données est indissociable des usages possibles à partir des données récoltées. Pour améliorer la qualité des données vis-à-vis de ces usages, il est nécessaire de mener en continu des études dédiées à ce sujet (10, 48, 49). De nombreux biais existent dans le processus de collecte des données des EDS. Ceux-ci sont liés à la couverture géographique de l'hôpital, aux changements dans les SI, aux différents processus existants au sein d'un même établissement. Le travail de mise en qualité permet de détecter, documenter et corriger ces biais. Cependant, c'est un sujet d'étude à lui seul encore trop peu valorisé comme tel, alors qu'il est précieux pour tous les projets menés sur EDSH. Du fait de leur culture en informatique médicale, de nombreuses équipes EDSH effectuent généralement ce travail de qualité en se focalisant essentiellement sur des questions médicales. Cette approche permet de cibler les sujets d'études les plus adaptés aux données disponibles. En limitant le champ à un sous-ensemble de données, cela facilite également le travail de mise en qualité. En revanche, des investissements supplémentaires dépassant le champ de l'étude initiale sont nécessaires afin de documenter les conclusions obtenues en termes de qualité et en faire bénéficier les études ultérieures. Ce sous-investissement à long terme sur le sujet de la qualité est aussi la conséquence d'un financement par projet. Des personnes spécialisées sur la qualité de la donnée dans chaque EDSH pourraient mener en permanence ce type d'études.

### Des premiers outils et processus de suivi de la qualité encore inégalement utilisés

Des processus réguliers de vérification de la cohérence avec un échantillon aléatoire de patients et une vérification manuelle devraient être mis en place et publiés, comme cela l'est déjà dans certains centres. Parmi les bonnes pratiques à généraliser, on peut également citer les alertes automatiques pour les utilisateurs sur leur périmètre de données, pointant les dysfonctionnements découverts par des scripts réguliers. Des tableaux de bord avec les caractéristiques agrégées des données peuvent être publiés sur les sites institutionnels (ex. : à l'institut Curie). Des études de qualité historique comme celle publiée par l'HEGP (49) sont également précieuses afin de servir de référence par la suite aux chercheurs travaillant sur l'EDS. Ces dispositifs aident à démystifier ce que contiennent les EDSH et améliorent la qualité des projets.

Certains outils spécialisés existent pour faciliter la création des flux, le traçage des changements, le suivi de la qualité et la documentation. Parmi ces logiciels d'*Extract Transform Load* (ETL), la préférence devrait être donnée à ceux mettant l'accent sur la documentation des flux et leur diffusion publique. Les outils *open source* du réseau de recherche OHDSI sont particulièrement adaptés aux recherches observationnelles[26]. Ils sont déjà utilisés par tous les EDSH participant au programme EHDEN.

[26] ohdsi.org/software-tools/

**La documentation technique des EDSH est encore à ses débuts**

Il est nécessaire de publier en accès libre (*open source*) les codes de recherche afin d'assurer une recherche rétrospective de qualité (10). Ce besoin de transparence concerne toute la chaîne de transformation de la donnée depuis les SI sources jusqu'aux *datamarts* utilisés pour les études, en passant par les schémas de données. Disposer d'une telle documentation permet également d'accélérer le prototypage, le développement et la mutualisation de nouveaux outils en découplant la donnée (non partageable) du code informatique. Les recherches récentes en analyse de données ont montré que des biais innombrables peuvent se cacher dans les jeux de données d'entraînement (50, 51). La publication ouverte des schémas de données est considérée comme un prérequis indispensable pour tous les usages de data science et d'intelligence artificielle (50).

Les exigences de sécurité et de garantie de protection de la vie privée n'entrent aucunement en contradiction avec ce principe de transparence. En revanche, les EDSH rencontrés soulignent la difficulté d'ajouter cette mission de documentation à leurs missions. Ceux accompagnés par le projet EHDEN ont salué les outils et les méthodes d'OHDSI obligeant la documentation précise des flux de données. En s'inspirant des *model cards* (52), *dataset cards* (50)[27] et du guide de publication d'un jeu de données sur datagouv.fr[28], il serait intéressant de définir une carte EDS, c'est-à-dire des spécifications techniques (métadonnées) pour la documentation des principaux flux de données.

## 4.2. Perspectives pour la HAS

### 4.2.1. Un écosystème en pleine expansion, à la recherche de pérennité

Les entretiens ont confirmé la richesse des données présentes dans les EDSH hospitaliers français. Des efforts de longue date sont menés dans les grands centres hospitaliers afin d'agréger les données sous une forme homogène et de les qualifier pour la recherche, le pilotage et le soin. Les formes de gouvernance se stabilisent et permettent d'envisager l'élaboration de cadres d'exploitation pour les instances de régulation.

En revanche, l'hétérogénéité actuelle des structures de données, des méthodes de mise en qualité ainsi que le manque de documentation publique sur les transformations effectuées depuis les sources de recueil ne permettent pas, à ce jour, une utilisation pour des études, recherches ou évaluations à l'échelle nationale. Le développement sur projet, le manque de financements pérennes, d'instances d'échange et de collaboration technique au niveau national ne favorisent pas la mise en place de solutions homogènes et interopérables sur le territoire.

Dans les prochaines années, cette situation est amenée à fortement évoluer grâce à la consolidation de l'écosystème des EDS, au travers de dynamiques de collaboration inter-CHU et l'appel à projets du ministère de la Santé et de la Prévention, opéré par la Plateforme des Données de Santé.

[27] Ces cartes de jeux de données sont très utilisées en traitement du langage, https://huggingface.co/docs/datasets/v1.12.0/dataset_card.html

[28] https://doc.data.gouv.fr/jeux-de-donnees/publier-un-jeu-de-donnees/

### 4.2.2. Lancement de projets expérimentaux pour éprouver le potentiel des EDSH pour les missions de la HAS

Les échanges nous ont permis de cibler deux opportunités de collaboration pour lesquelles les EDSH hospitaliers semblent dès aujourd'hui prometteurs.

- **La contextualisation de l'utilisation des produits de santé :** certains actes ou médicaments sont majoritairement réalisés et administrés en milieu hospitalier. Pour ce type de soins, il est important pour la HAS de mieux connaître leur cadre d'utilisation, les pathologies concernées et les caractéristiques des patients soignés. Des études descriptives à partir de plusieurs EDSH français permettraient de mieux comprendre leur place dans les soins de routine. Cette connaissance est précieuse afin d'évaluer ces actes en vue de leur remboursement dans le cadre du droit commun.

  C'est le cas des actes de séquençage en oncologie. Ceux-ci sont collectés dans le Référentiel des actes innovants hors nomenclatures (RIHN). Ces actes ne sont collectés à l'échelle nationale que par un décompte à très grosse maille (trois codes seulement) agrégé par établissement. La plupart de ces données sont récoltées dans les logiciels de la biologie de façon structurée. L'identification des actes précis est plus simple que le repérage d'information à partir de données textuelles, pour lesquelles un travail long et complexe d'extraction et de normalisation d'information est nécessaire. Cette structure préexistante dans les SIH permet d'envisager des projets courts s'inscrivant dans les temporalités de la HAS. L'objectif du projet sera d'obtenir la description de l'activité des actes de séquençage haut débit (*Next Generation Sequencing*) inscrits au RIHN dans le cadre de la cancérologie.

  Précisément, il s'agira d'obtenir des quantités d'actes effectués pour chaque variant génétique recherché et les principales caractéristiques des patients (diagnostics codés, démographie). Ces résultats seront confrontés à ceux obtenus par le biais de questionnaires envoyés aux conseils nationaux professionnels, plateformes d'oncogénétique, sociétés savantes et autres groupements de professionnels d'intérêt.

- **Le développement de proxys de qualité et de sécurité des soins :** différents indicateurs de qualité et de sécurité des soins (IQSS) sont développés et validés par la HAS en lien avec les professionnels de santé, les patients et usagers[29]. Ils sont mis à disposition des professionnels de santé pour l'amélioration de la qualité et de la sécurité des soins en établissements de santé. Une large part de ces IQSS sont mesurés annuellement dans les hôpitaux par un retour manuel aux dossiers de patients tirés au sort. Cela représente une charge de travail importante pour les établissements.

  L'opportunité de recueillir une partie de ces indicateurs à partir des dossiers patients informatisés semble prometteuse à divers titres. Cela permettrait d'automatiser tout ou partie du recueil, et donc de moins solliciter les équipes en établissement de santé. Cela ouvrirait également de nouvelles possibilités, telles qu'un recueil exhaustif sur un établissement donné ou une fréquence d'actualisation plus élevée, permettant des retours d'information réguliers sur leur pratique aux professionnels de santé.

  Les premiers échanges avec les EDSH semblent indiquer que certains indicateurs peuvent être automatisés, mais en passant par des approximations, lorsqu'ils mobilisent des données facilement structurables dans les EDSH. Il s'agit pour la HAS de comprendre ce qui peut être automatisé, à quel coût. De plus, les EDSH ne couvrent pas tous les hôpitaux, donc le passage à l'échelle nationale n'est pas

[29] https://www.has-sante.fr/jcms/c_970481/fr/indicateurs-de-qualite-et-de-securite-des-soins-en-etablissements-de-sante

envisagé tel quel. Un enjeu majeur sera l'identification et la définition de proxys de qualité satisfaisants. En effet, il n'existe pas, dans les EDSH, de variable mesurant directement la qualité ou la sécurité des soins. En revanche, on peut espérer trouver un ensemble de variables, qui, correctement agrégées, peuvent former un indicateur reflétant fidèlement les notions de qualité ou de sécurité des soins.

La HAS envisage ces deux cas d'usages comme des projets pilotes permettant de tester la mobilisation des EDSH pour ses besoins systématiques. Selon le principe de co-construction des entrepôts et de leurs usages, la HAS considère que les données de soins des hôpitaux ne montreront leur vrai potentiel qu'une fois leur exploitation confrontée à des projets concrets.

En 2023, la HAS va définir les cadres de collaboration et préciser les modalités techniques de ces projets, en lien avec les CHU intéressés.

## 4.3. Limites de l'analyse

Les entretiens ont été menés de façon semi-dirigée dans un laps de temps limité. Certaines thématiques ont donc été couvertes plus rapidement et seuls les éléments explicitement mentionnés par les participants ont pu être relevés.

L'existence inégale des portails d'études introduit un biais dans le relevé des types d'études menées sur les EDSH. Ceux ayant un portail de transparence ont déjà une certaine maturité dans les cas d'usages.

Nous avons interrogé uniquement 17 des 32 CHU, parmi ceux ayant les dynamiques d'entrepôt les plus avancées. Nous compléterons ce panorama en interrogeant les CHU restants qui ont une démarche prospective d'EDSH et partagerons les résultats dans une publication séparée. De plus, avec seulement 1 CLCC et 4 groupes hospitaliers dont 2 établissements de santé privés d'intérêt collectif, nous avons peu couvert les autres structures de soins. Il existe des initiatives semblables en ville, dans des groupes hospitaliers de plus petite taille et au sein d'entreprises privées. De plus, nous n'avons pas couvert les entrepôts spécialisés sur des thématiques cliniques (cancers, VIH, maladies rares).

Nous avons peu abordé le modèle de financement des EDSH. Nous avons collecté peu d'informations concernant les budgets pour la mise en place et le fonctionnement des EDSH, ou les budgets nécessaires à la mise en œuvre des projets sur EDSH. Cette question mériterait un travail approfondi, dans la durée. Il faudrait aussi trouver des moyens de mesurer les bénéfices des EDSH, qui – comme pour toute infrastructure –sont souvent indirects et sur le long terme. Enfin, dans un contexte extrêmement tendu concernant les compétences nécessaires au sein des équipes EDSH, il faudrait identifier les mécanismes financiers favorisant la collaboration entre institutions plutôt que la concurrence.

Enfin, nous n'avons pas étudié la question de la responsabilité du contenu présent dans les EDSH. L'office américain des technologies de l'information en santé distingue le DPI *Electronic Medical Record* (EMR) de la collection systématique des données de santé du patient dans un *Electronic Health Record* (EHR)[30]. À un EMR donné correspond une personne morale, opérateur de soins et responsable de traitement. Les EDSH rassemblent ces données au sein d'un *Electronic Health Record* (EHR). En pratique, ceux-ci ont pour responsables de traitement des opérateurs qui n'ont pas de responsabilité de soins. Ils portent donc moins de responsabilité quant à la qualité médicale ou même structurelle des données qu'ils exposent. Il serait pertinent d'interroger le lien entre cette responsabilité et la qualité des données présentes dans les SIH ou les EDSH.

[30] https://www.healthit.gov/faq/what-are-differences-between-electronic-medical-records-electronic-health-records-and-personal

# Conclusion

L'écosystème des entrepôts de données de santé est en pleine construction. Il bénéficie actuellement d'une accélération grâce à des financements nationaux, la multiplication d'acteurs industriels spécialisés en données de santé et le début d'une réflexion supranationale à propos de l'espace européen de données de santé.

Lors de cette phase de développement des EDSH, la HAS a identifié plusieurs éléments importants afin de mieux valoriser le potentiel des données de santé.

La HAS recommande la constitution et la pérennisation d'équipes entrepôts multidisciplinaires en mesure d'opérer l'EDSH et d'accompagner les différents projets. Cette équipe devrait être le point de contact privilégié concernant les sujets d'exploitation de la donnée en collaborant avec les autres directions métiers impliquées : DSI, DIM, DRCI, directions cliniques.

La constitution d'une gouvernance à trois niveaux est un autre chantier prioritaire que préconise la HAS. Les usages multicentres nécessitent de créer des consensus sur les questions d'interopérabilité, de schéma de données, de nomenclatures et de mise en qualité. Des coordinations interrégionales et nationale permettraient de créer des groupes de travail thématiques, afin d'impulser une dynamique de coopération et de mutualisation.

La HAS recommande également la constitution d'un socle commun de données, avec des métadonnées précises permettant de cartographier les données intégrées, afin de qualifier les usages à développer dès aujourd'hui à partir des EDSH. Plus largement, la documentation *open source* des flux de données et des transformations effectuées pour la mise en qualité nécessiterait plus d'incitations afin de libérer le potentiel d'innovation pour tous les réutilisateurs de données de santé.

Enfin, la question de l'élargissement du périmètre des données, au-delà du domaine purement hospitalier, doit être posée. De nombreux facteurs de risque et des données de suivi des patients sont absents des EDSH, mais cruciaux afin de comprendre les pathologies. Combiner les données de ville et les données hospitalières permettrait d'avoir finalement une vision complète sur la prise en charge des patients. La HAS préconise qu'une réflexion pour la systématisation des appariements entre les données des EDSH et les données de facturation soit engagée, quitte à ne fournir le fruit de cet appariement qu'aux projets le nécessitant.

# Table des annexes

# Annexe 1. Stratégie de recherche documentaire

**Méthode**

La recherche documentaire dans les bases de données bibliographiques Medline, Embase et Emcare a porté sur la période janvier 2016-décembre 2021 et a été limitée aux publications en langue française et anglaise.

La stratégie de recherche a visé les publications des équipes françaises.

Les sources internet suivantes ont par ailleurs été interrogées :

- sites internet dédiés au numérique en santé ;
- sites internet d'hôpitaux ou de fédérations hospitalières ;
- sites internet d'universités ;
- sites internet dédiés à la législation.

**Bases de données bibliographiques**

La stratégie de recherche dans les bases de données bibliographiques est construite en utilisant, pour chaque sujet, soit des termes issus de thésaurus (descripteurs), soit des termes libres (du titre ou du résumé). Ils sont combinés avec les termes décrivant les types d'études.

Le tableau suivant présente la stratégie de recherche dans les bases de données Medline, Embase et Emcare.

| **Type d'étude/Sujet** | | **Période de recherche** | **Nombre de références trouvées** |
|---|---|---|---|
| | Termes utilisés | | |
| **Entrepôts de données cliniques**<br>***Recommandations***<br>**France** | | | |
| | | Janv. 2016<br>Déc. 2021 | 11 |
| Étape 1 | (clinical data warehous* OR clinician data warehous* OR medical data warehous* OR health data warehous* OR clinical datawarehous* OR clinician datawarehous* OR medical datawarehous* OR health datawarehous* OR clinical data reposiser* OR clinician data reposiser* OR medical data reposiser* OR health data reposiser* OR clinical data mart* OR clinician data mart* OR medical data mart* OR health data mart* OR electronic clinical record* OR electronic medical record* OR electronic health record*)/titre OR (Electronic Health Records OR Electronic Health Record OR Hospital Information Systems OR Hospital Information System | | |

| | | | |
|---|---|---|---|
| | OR Data Warehousing OR Data Warehouse OR Electronic Medical Record System OR Clinical Data Repository)/descripteur | | |
| AND | | | |
| Étape 2 | (francais* OR french* OR France)/ti,ab OR France/de OR (francais* OR français*)nom_revue OR (français* OR France)/affiliation OR (France)/localisation OR (France)/localisation_sujet OR (France)/pays OR (France)/lieu_de_travail | | |
| AND | | | |
| Étape 3 | (consensus OR guidance OR guideline* OR guide OR position paper OR recommendation* OR statement*)/titre OR (health planning guidelines OR consensus development OR Practice Guideline)/descripteur OR (consensus development conference OR consensus development conference, NIH OR guideline OR practice guideline OR Government Publication)/type | | |
| **Entrepôts de données cliniques**<br>***Revues systématiques***<br>**France** | | | |
| | | Janv. 2016<br>Déc. 2021 | 4 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 4 | (systematic* overview* OR systematic* research* OR systematic* review* OR systematic* search*)/titre OR systematic review/descripteur OR systematic review/type OR (cochrane database syst rev OR Health Technol Assess)/revue | | |
| **Entrepôts de données cliniques**<br>***Autres revues***<br>**France** | | | |
| | | Janv. 2016<br>Déc. 2021 | 34 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 5 | review/titre OR review/type | | |
| **Entrepôts de données cliniques**<br>***Conception***<br>**France** | | | |

| | | | |
|---|---|---|---|
| | | Janv. 2016<br>Déc. 2021 | 33 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 6 | (design OR designs OR designing OR develop OR develops OR developing OR developement OR development OR construct OR constructs OR constructing OR construction OR create OR creates OR creating OR creation OR build OR builds OR building OR conception OR architecture OR landmark* OR project* OR pilot*)/titre OR (Program Development OR Equipment Design OR Data Science OR pilot study)/descripteur | | |

**Entrepôts de données cliniques**

***Implémentation***

**France**

| | | | |
|---|---|---|---|
| | | Janv. 2016<br>Déc. 2021 | 85 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 7 | (launch OR launches OR launching OR implement OR implements OR implementation OR implementing OR deploy OR deploys OR deploying OR validate OR validates OR validation OR operationalize OR operationalizes OR operationalization OR introduce OR introduces OR introduction OR evolution)/titre OR (Implementation Science OR Health Plan Implementation OR Program Evaluation OR Clinical Coding OR Coding OR Data Aggregation OR Data Integration OR Data Integrity OR Data Interoperability OR Health Information Interoperability OR Systems Integration OR Medical Record Linkage OR Electronic Data Interchange OR Database Management System OR Reporting And Data System OR Artificial Intelligence OR Natural Language Processing OR Data Mining OR Data Visualization OR Learning Health System OR Machine Learning OR Pattern Recognition, Automated OR Automated Pattern Recognition OR Expert System OR Expert Systems)/descripteur | | |

**Entrepôts de données cliniques**

***Études de pratiques***

**France**

| | | | |
|---|---|---|---|
| | | Janv. 2016<br>Déc. 2021 | 49 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |

| | | | |
|---|---|---|---|
| Étape 8 | (example* OR experience* OR experiment* OR practices OR initiative* OR pattern* OR trend* OR survey* OR scope OR governance OR challenge* OR barrier* OR pitfall* OR facilitator* OR attitude* OR behavior* OR behaviour* OR discourse* OR interview* OR stories OR story)/titre OR (Electronic Health Records -- statistics & numerical data OR Hospital Information Systems -- statistics & numerical data OR Data Warehousing -- statistics & numerical data)/descripteur | | |
| **Entrepôts de données cliniques**<br>***Impact sur la qualité des soins***<br>**France** | | | |
| | | Janv. 2016<br>Déc. 2021 | 29 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 9 | ((quality OR performance OR competence) AND (evaluat* OR indicator* OR criteria OR criterium OR measur* OR assess* OR improv* OR standard* OR control* OR management))/titre OR TI(benefit OR benefits OR impact)/titre OR (Quality Improvement OR quality improvement study OR clinical effectiveness OR Quality of Health Care OR health care quality OR Outcome and Process Assessment, Health Care OR Process Assessment, Health Care OR Health Care Evaluation Mechanisms OR Quality Indicators, Health Care OR performance measurement system OR Clinical Audit OR Quality Control OR Quality Assurance, Health Care OR Total Quality Management OR Practice Patterns, Nurses' OR Practice Patterns, Dentists' OR Practice Patterns, Physicians' OR Practice Patterns, Pharmacists' OR protocol compliance OR Guideline Adherence)/descripteur | | |
| **Entrepôts de données cliniques**<br>***Aspects économiques***<br>**France** | | | |
| | | Janv. 2016<br>Déc. 2021 | 9 |
| Étape 1 AND Étape 2 | | | |
| AND | | | |
| Étape 10 | (cost OR costs OR costing OR economic OR economics OR economical OR economy OR expensive OR inexpensive OR finance OR financial OR financing OR price OR prices OR pricing OR spend OR spending OR spent OR efficient OR efficience OR sustainable OR sustainability OR business OR market)/titre OR (Electronic Health Records -- economics OR Hospital Information Systems -- economics OR Data Warehousing -- economics OR | | |

| | | | |
|---|---|---|---|
| | electronic health record -- device economics OR hospital information system -- device economics OR data warehouse -- device economics)/descripteur OR (Budgets OR Cost-Benefit Analysis OR Costs and Cost Analysis OR Models, Econometric OR Models, Economic OR budget OR cost benefit analysis OR econometric model OR economic model OR economic aspect OR market)/descripteur | | |

À la suite de la recherche ci-dessus, un classement des auteurs selon le nombre de leurs publications identifiées dans le corpus ainsi obtenu a également été réalisé afin d'identifier les principaux auteurs français sur ce sujet.

**Sites internet consultés**

Dans le cadre de cette évaluation, les sites suivants ont été consultés :

Plateforme des données de santé

https://www.health-data-hub.fr/

BercyNumérique

https://www.bercynumerique.finances.gouv.fr/

Etalab

https://www.etalab.gouv.fr/accompagnement-logiciels-libres/

data.gouv.fr

https://doc.data.gouv.fr/jeux-de-donnees/publier-un-jeu-de-donnees/

Ministère de la Santé et de la Prévention

https://solidarites-sante.gouv.fr/

Haute Autorité de santé

https://www.has-sante.fr/

Légifrance

https://www.legifrance.gouv.fr/

Assistance publique-Hôpitaux de Paris (AP-HP) – Entrepôt de données de santé

https://eds.aphp.fr/

CHRU Hôpitaux de Tours

https://www.chu-tours.fr/recherche-et-innovation/recherche-professionnels/recherche-clinique-et-translationnelle/monter-un-projet-de-recherche/

CHU Angers – Centre de données cliniques

https://www.chu-angers.fr/recherche-et-innovation-sante/structures-d-appui-a-la-recherche/centre-de-donnees-cliniques/

CHU Brest – Entrepôt de données de santé

https://www.chu-brest.fr/fr/entrepot-donnees-sante

CHU Grenoble Alpes – Entrepôt de données de santé (EDS)

https://www.chu-grenoble.fr/content/entrepot-de-donnees-de-sante-eds

CHU Lille

https://www.chu-lille.fr/rgpd-recherche

CHU Nantes

https://www.chu-nantes.fr/liste-de-recherches-sur-donnees

CHU Reims – Entrepôt de données de santé

http://www.iias.fr/p/Entrepot-de-Donnees-de-Sante

CHU Rennes – Centre de données cliniques – Entrepôt de données hospitalier eHop

https://www.chu-rennes.fr/centre-donnees-cliniques.html

CHU Rouen Normandie – Entrepôt de données de santé normand (EDSaN)

https://edsan.chu-rouen.fr/edsan/acces-aux-donnees/procedures/

Hôpital Foch

https://www.hopital-foch.com/patients-familles/recherche/les-etudes-sur-donnees-realisees-a-foch/liste-des-etudes/

Hôpitaux universitaires Grand-Ouest (HUGO)

https://www.chu-hugo.fr

Hospices civils de Lyon (HCL)

https://myhcl.sante-ra.fr/Espacepublic/ListeEtudesSurDonneesHCL.aspx

Institut Curie – *Data Factory*

https://github.com/curie-data-factory/health-data-metrics/

Fédération de l'hospitalisation privée

https://www.fhp.fr/

Université de Rennes 1 – Centre de données cliniques

https://centrededonneescliniques.univ-rennes1.fr/

imagine-bdd

https://github.com/imagine-bdd/DRWH/

InterHop

https://interhop.org/

Codoc

https://codoc.co/

Arkhn

https://arkhn.org/

Cegedim

https://www.cegedim-health-data.com/

OpenSAFELY

https://www.opensafely.org/

*European Commission* – joinup

https://joinup.ec.europa.eu/

*European Health Data & Evidence Network* (EHDEN)

https://www.ehden.eu/

*Office of the National Coordinator for Health Information Technology* (ONC) – HealthIT.gov

https://www.healthit.gov/faq/what-are-differences-between-electronic-medical-records-electronic-health-records-and-personal

*Stanford Medicine – Research IT – Technology & digital solutions*

https://med.stanford.edu/researchit/news/CDW-reimagined.html

*US Food and Drug Administration – Real World Evidence*

https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

*Observational Health Data Sciences and Informatics* (OHDSI)

https://ohdsi.org/

## Annexe 2. Répartition des acteurs interrogés par source de repérage

Les acteurs ont été repérés par les moyens suivants :

- leur rôle institutionnel : Plateforme des Données de Santé , DGOS, Inca, Fédération de l'hospitalisation privée ;
- Légifrance : recherche des autorisations CNIL pour un entrepôt de données de santé à la suite du référentiel correspondant (délibération n° 2021-118 du 7 octobre 2021). Recherche effectuée grâce aux mots clés entrepôts de données de santé sur le site de Légifrance en février 2022 : AP-HP, APHM, CHU de Rouen, CHU d'Angers, CHU de Lille, CHU Grenoble Alpes, CHU de Rennes, institut Curie, CHU de Nantes ;
- les premiers entretiens menés ayant permis d'identifier de nouveaux acteurs à interroger : Arkhn, Codoc, CHU de Brest, CHU de Nancy, CHU d'Amiens, CHU de Bordeaux ;
- à travers le réseau d'ingénieurs en données de santé InterHop[31] : Foch, CHU de Toulouse, hôpital européen Georges Pompidou, InterHop ;
- des recherches bibliographiques sur les bases Medline, Embase, Emcare visant à identifier les EDSH déjà mis en place en France (termes exacts de la recherche en Annexe 5.3) : Montpellier, institut Imagine-Necker, HCL, CHU de Tours, groupe hospitalier Paris Saint-Joseph ;
- le travail de référencement déjà effectué par le ministère de la Santé (DGOS, bureau PF5) pour son appel à projets : CHU de Poitiers.

[31] https://interhop.org/

## Annexe 3. Acteurs interrogés, liste exhaustive

| Entrepôt/organisation | Équipes |
|---|---|
| EDS_AMIENS | DIM : 1, DSI : 1 |
| EDS_ANGERS | Direction des données : 1 |
| EDS_APHM | Clinicien : 1, Entrepôt : 2 |
| EDS_APHP | DSI : 5, Entrepôt : 4 |
| EDS_BORDEAUX | Entrepôt : 1, Inserm : 1, santé publique : 2 |
| EDS_BREST | DIM : 1, Entrepôt : 1 |
| EDS_CURIE | Direction des données : 1, Entrepôt : 1 |
| EDS_EDSAN | DIM : 1, Entrepôt : 2 |
| EDS_FOCH | DRCI : 2, DSI : 1 |
| EDS_HCL | Clinicien : 1, DRCI : 1, DSI : 1, direction des données : 1 |
| EDS_HEGP | BDNMR : 1, DIM : 2, Inserm : 1, bio-informatique : 1 |
| EDS_IMAGINE | Entrepôt : 1 |
| EDS_INCA | Direction des données : 1 |
| EDS_INCLUDE_LILLE | Administration : 2, Entrepôt : 3, santé publique : 2 |
| EDS_MONTPELLIER | DIM : 1, direction des données : 2, santé publique : 1 |
| EDS_NANCY | Entrepôt : 2, santé publique : 2 |
| EDS_NANTES | Entrepôt : 2 |
| EDS_POITIERS | DRCI : 1, DSI : 2 |
| EDS_PREDIMED_CHUGA | Entrepôt : 3, santé publique : 2 |
| EDS_RENNES | Entrepôt : 2, santé publique : 2 |
| EDS_SAINT_JOSEPH | Santé publique : 1 |
| EDS_TOULOUSE | Entrepôt : 1 |
| EDS_TOURS | Entrepôt : 1 |
| Arkhn | Startup : 1 |
| Codoc | Startup : 1 |
| DGOS | Administration : 1 |
| Plateforme des Données de Santé | Administration : 1 |
| InterHop | Clinicien : 1 |

## Annexe 4. Formulaire d'entretien

| Thématique | Questions |
|---|---|
| Initiation et Construction de l'Entrepôt de Donnée de Santé | Comment est né l'initiative, quand, quelle(s) équipe(s) impliquées dans la construction ? Un entrepôt pour répondre à quels besoins initiaux ? |
| | Quelle a été/est l'articulation entre les équipe(s) d'informatique médicale / ingénieur(s) / DRCI / et les équipe(s) usagers, les biostatistiques ? |
| | Gouvernance : Quelle organisation des équipes pour la constitution et maintenance de l'entrepôt, l'accès aux données, les équipes projets ? |
| | Quels sont les types de données présentes dans l'entrepôt parmi la liste non-exhaustive suivante :<br>facturations (PMSI), autres données administratives, autres actes, interventions et diagnostics structurés, mesures de biologies structurées,<br>traitements médicamenteux structurés, urgences, réanimation, anesthésie, textes (courriers, CR), imagerie, anatomopathologie, séquençage. |
| | Quelles sont les données à caractère médico-social/social, notamment provenant d'établissements sociaux et médico-sociaux ? |
| État des lieux actuel - Projets menés | Qui sont les principaux utilisateurs ? Pour quels besoins (recherche, amélioration de la qualité des soins, pilotage, clinique) ?<br>Quelle(s) aire(s) thérapeutiques ? |
| | Quels sont les grands types de projets parmi la liste non-exhaustive suivante :<br>création de cohorte, épidémiologie descriptive, épidémiologie analytique (comparative) avec/sans randomisation, pilotage et tableaux de bords,<br>alertes et indicateurs, inclusion dans des essais cliniques. |
| | Quel est le nombre de projets terminés / entamés / projetés ? |
| | Quels sont les outils et méthodes utilisés pour ces projets ? Outil de constitution de cohorte, formats standards de données, NLPs, … |
| | Quelle valorisation de l'entrepôt ? |
| | Quels liens avec des sources externes (données de ville, HDH, cohortes) ? |
| Opportunités et obstacles | Quelles sont les principales difficultés rencontrées lors des projets menés sur l'entrepôt de données ? |
| | Y-a-t-il des thématiques qui mériteraient plus d'incitation de la part de la HAS ? |
| | Quelles compétences sont indispensables ? Manque-t-il des compétences, des ressources techniques ? |
| Critères de qualité pour la recherche observationnelle | Couverture : Comment est-elle contrôlée ? Sur le plan géographique/ par service ? Sur le plan temporel ? Par quels moyens ? |
| | Nettoyage : Comment se fait la gestion des duplicata patients et de l'alignement des sources ? |
| | Réseau de base de données : Est-ce que l'entrepôt appartient à un réseau de base de données de santé ?<br>Lien vers les études qui en sont sorties. |
| | Qualité de la donnée : Est-ce qu'il existe des rapports automatiques sur la qualité des données ? Fréquence, design,<br>code et documentation accessible ? Présence de personnel dédié voire d'une équipe pour vérifier la qualité des données en continu,<br>et effectuer des contrôles qualité des données sur la base centrale, sur les bases d'étude ? |
| | Cycle de vie de la donnée : Y a t-il un document de référence sur les différentes étapes du cycle de vie de la donnée ?<br>Comment ce document est il tenu à jour vis-à-vis des évolutions constantes de l'entrepôt ? Sous quelle forme ?<br>Quel est le mode de gestion, d'accès, d'actualisation, de correction de cette documentation ? Description précises des champs intégrés ? |
| | Procédure d'harmonisation : Quels sont les structures / formats de données et les système de codages utilisés ?<br>(eHop, I2B2, Omop, Dr Warehouse, FHIR, autre ?) |
| | Apprentissage automatique : Si des systèmes d'apprentissage automatique sont utilisé (par exemple pour extraire et structurer de l'information),<br>y-a-t-il une documentation spécifique sur leurs performances ? En ce qui concerne le codage manuel (par exemple labelling),<br>le guide de codage existe-t-il ? Est-ce qu'une mesure de la cohérence inter-codeurs a été menée ? |
| | Dé-identification : Eléments sur la dé-identification si applicable, métriques de performance |
| | Phénotypes construits : Existe-t-il des définitions opérationnelles des populations cibles (study cohorts) et comment celles-ci sont elles<br>confrontés aux définitions conceptuelles ie. métiers et scientifiques ? Une étude des FPR/TPR par rapport à un standard de référence existe-t-il ?<br>Ces définitions sont elles rendues publiques soit avec les résultats d'étude, soit dans la documentation de l'entrepôt ? |
| | Transparence: Les études sont elles enregistrées sur un portail dédié ou pré-existant (épidémio-France, encepp (EU), clinicaltrials.gov (US)) ?<br>Les codes d'études sont ils rendus accessibles comme pour opensafely ? Les publications sont-elles accessibles en open-access, une fois les études terminées ? |
| | Multidisciplinarité : Les équipes projets sont-elles multi-disciplinaires ? Spécification des participations pour chaque partie de<br>l'analyse depuis la collection des données depuis le SI source. |
| Sujets d'intérêt pour la HAS | Direction de la Qualité : IQSS : coordination (evaluation du patient pour la sortie, prise de contact du patient à J+1), qualité de la lettre de liaison),<br>prise en charge (eligibilité à l'intervention en chirambu, prise en charge de la douleur) |
| | Direction de l'Evaluation : Activité de biologie hospitalière (description), effets indésirables associés aux actes,<br>études post-inscription (actes, accès précoces oncologie). Evaluation des actes : ex. actes de biologie + imagerie réalisés à l'hôpital,<br>tests génétiques en oncologie et maladies rares. |
| Echange libre | |

## Annexe 5. Spécialités des promoteurs d'étude sur données

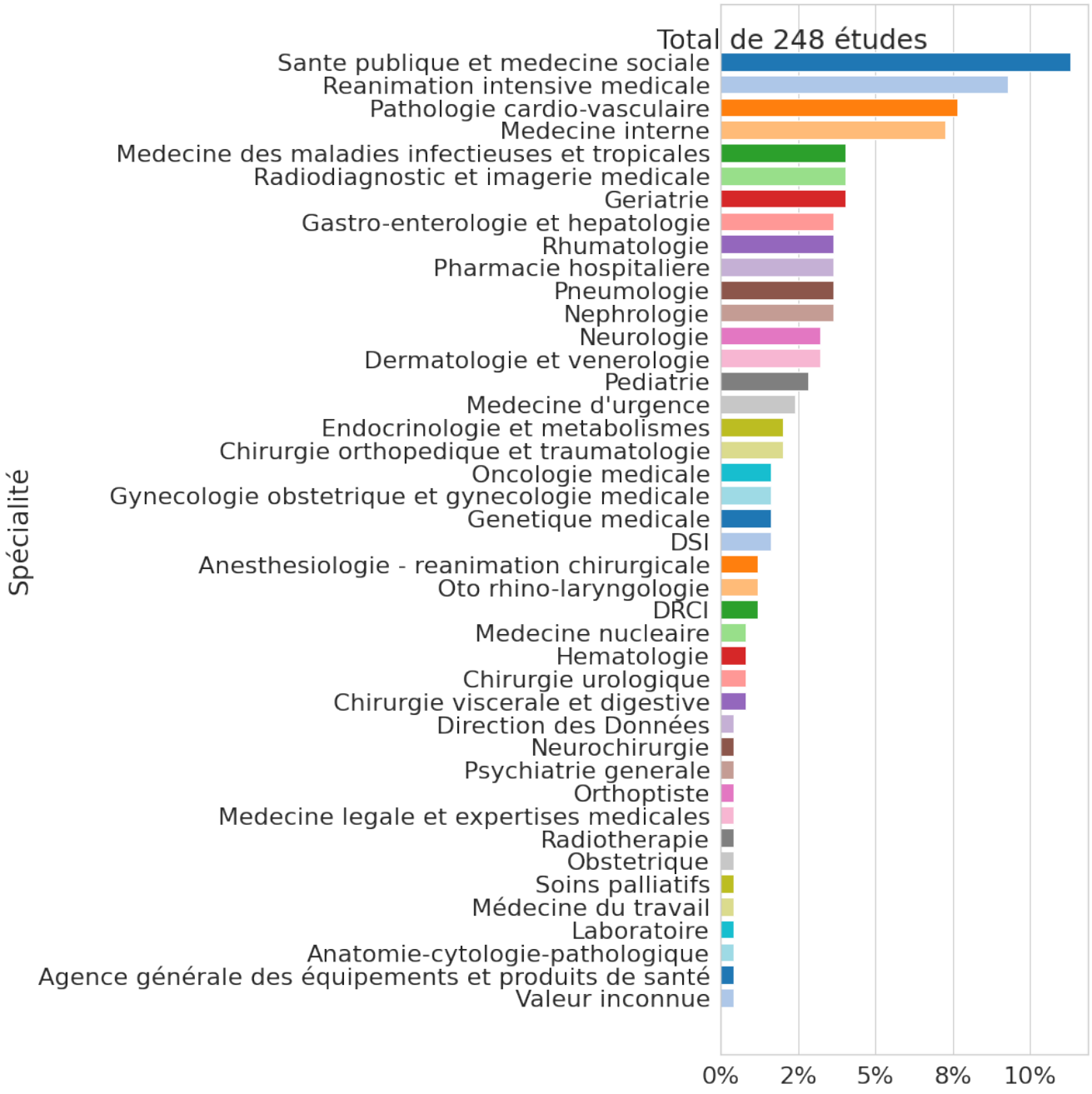


**Figure 7 – Répartition des études sur EDSH disponibles sur les portails de transparence par spécialité du promoteur de l'étude**

## Annexe 6. Tables de résultats

Les tables de données ayant servi à produire les figures de la partie résultats sont disponibles à l'adresse suivante : https://gitlab.has-sante.fr/has-sante/public/rapport_edsh.

La table des **invités** concerne les individus interrogés, les dates d'entretien, les postes et l'appartenance à une équipe spécifique.

La table des **entrepôts** collige des informations sur les EDSH.

La table des **études** est le référencement des études renseignées sur 10 portails d'études en cours (ou terminées si celles-ci sont disponibles) disponibles en accès libre.

### Colonnes renseignées dans chaque table

**- Table invités :** nom, date interview, organisation, poste, equipe_categorie, entrepot_id

**- Table entrepôts :** entrepot_id, organisation, url_portail_etudes_en_cours, n_etp_entrepot, n_patients, n_etudes_en_cours, common_data_model, url_procedures_access, datalab, doc_cyle_de_vie_donnee, equipe_qualite, date_debut, date_debut_cnil, données, deidentification, valorisation_text_col, DPI, private_actors, cycle_eds_invites, id_fi_ej, id_fi

**- Table études :** titre_etude, titre_etude_long, type_etude, methode_etude, domaine_medical, covid, objectif, date_collecte, entrepot_id, source

# Références bibliographiques


1. Food and Drug Administration. Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products : FDA; 2021.
https://www.fda.gov/media/152503/download

2. Haute Autorité de Santé. Real-world studies for the assessment of medicinal products and medical devices. Saint-Denis la Plaine: HAS; 2021.
https://www.has-sante.fr/upload/docs/application/pdf/2021-06/real-world_studies_for_the_assessment_of_medicinal_products_and_medical_devices.pdf

3. National Institute for Health and Care Excellence, Kent S, Kincaid L, Manuj S, Rowark S, Duffield S*, et al.* NICE real-world evidence framework. London: NICE; 2022.
https://www.nice.org.uk/corporate/ecd9/resources/nice-realworld-evidence-framework-pdf-1124020816837

4. Institut national d'excellence en santé et en services sociaux, Plamondon G, Auclair Y, Dufort P, Beha S, Gonthier C*, et al.* Intégration des données et des preuves du contexte réel dans les évaluations en appui à la prise de décision dans le secteur des médicaments. Québec: INESSS; 2022.
https://www.inesss.qc.ca/fileadmin/doc/INESSS/Rapports/Medicaments/INESSS_Donnees_preuves_contexte_reel_EC.pdf

5. Haute Autorité de Santé. Cartographie des impacts organisationnels pour l'évaluation des technologies de santé. Guide méthodologique. Saint-Denis La Plaine: HAS; 2020.
https://www.has-sante.fr/upload/docs/application/pdf/2020-12/guide_methodologique_impacts_organisationnels.pdf

6. Flynn R, Plueschke K, Quinten C, Strassmann V, Duijnhoven RG, Gordillo-Marañon M*, et al.* Marketing authorization applications made to the european medicines agency in 2018–2019: What was the contribution of real-world evidence? Clin Pharmacol Ther 2022;111(1):90-7.
http://dx.doi.org/10.1002/cpt.2461

7. Arlett P, Kjær J, Broich K, Cooke E. Real-World Evidence in EU Medicines Regulation: Enabling Use and Establishing Value. Clin Pharmacol Ther 2022;111(1):21-3.
http://dx.doi.org/10.1002/cpt.2479

8. Food and Drug Administration. Real World Evidence Program. Silver Spring: FDA; 2018.
https://www.fda.gov/media/120060/download

9. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L*, et al.* Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique 2017;65:S149-S67.
http://dx.doi.org/10.1016/j.respe.2017.05.004

10. Kohane IS, Aronow BJ, Avillach P, Beaulieu-Jones BK, Bellazzi R, Bradford RL*, et al.* What every reader should know about studies using electronic health record data but may be afraid to ask. J Med Internet Res 2021;23(3):e22219.
http://dx.doi.org/10.2196/22219

11. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA*, et al.* BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. N Engl J Med 2021;384(15):1412-23.
http://dx.doi.org/10.1056/NEJMoa2101765

12. Gehring S, Eulenfeld R. German medical informatics initiative: Unlocking data for research and health care. Methods Inf Med 2018;57(S 01):e46-e9.
http://dx.doi.org/10.3414/me18-13-0001

13. Cuggia M, Garcelon N, Campillo-Gimenez B, Bernicot T, Laurent J-F, Garin E*, et al.* Roogle: an information retrieval engine for clinical data warehouse. Stud Health Technol Inform 2011;169:584-8.
http://dx.doi.org/10.3233/978-1-60750-806-9-584

14. Jannot A-S, Zapletal E, Avillach P, Mamzer M-F, Burgun A, Degoulet P. The Georges Pompidou University Hospital Clinical Data Warehouse: A 8-years follow-up experience. Int J Med Inform 2017;102:21-8.
http://dx.doi.org/10.1016/j.ijmedinf.2017.02.006

15. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F*, et al.* Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform 2017;73:51-61.
http://dx.doi.org/10.1016/j.jbi.2017.07.016

16. Wack M. Installation d'un entrepôt de données cliniques pour la recherche au CHRU de Nancy : déploiement technique, intégration et gouvernance des données [other]. : Université de Lorraine; 2017.
https://hal.univ-lorraine.fr/hal-01931928

17. Daniel C, Serre P, Orlova N, Bréant S, Paris N, Griffon N. Initializing a hospital-wide data quality program. The AP-HP experience. Comput Methods Programs Biomed 2018:104804.
http://dx.doi.org/10.1016/j.cmpb.2018.10.016

18. Malafaye N, Demoulin D, Mailhe P, Morell M, Pellecuer D, Dunoyer C. Mise en place et exploitation d'un entrepôt de données au département d'information médicale du CHU de Montpellier, France. Rev Epidemiol Sante Publique 2018;66:S26.
http://dx.doi.org/10.1016/j.respe.2018.01.055

19. Artemova S, Madiot P-E, Caporossi A, group P, Mossuz P, Moreau-Gaudry A. PREDIMED: Clinical Data Warehouse of Grenoble Alpes University Hospital. Stud Health Technol Inform 2019;264:1421-2.

http://dx.doi.org/10.3233/SHTI190464

20. Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a Semantic Health Data Warehouse in the Context of Clinical Trials: Development and Usability Study. JMIR Med Inform 2019;7(4):e13917.
http://dx.doi.org/10.2196/13917

21. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M-M*, et al.* Transforming French Electronic Health Records into the Observational Medical Outcome Partnership's Common Data Model: A Feasibility Study. Appl Clin Inform 2020;11(1):13-22.
http://dx.doi.org/10.1055/s-0039-3402754

22. Conan Y, Herbert J, Salpetrier C, Godillon L, Fourquet F, Dhalluin T*, et al.* Les entrepôts de données cliniques : un outil d'aide au pilotage de crise. Infect Dis Now 2021;51(5):S56.
http://dx.doi.org/10.1016/j.idnow.2021.06.119

23. Gunter TD, Terry NP. The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J Med Internet Res 2005;7(1):e3.
http://dx.doi.org/10.2196/jmir.7.1.e3

24. Hoerbst A, Ammenwerth E. Electronic Health Records: A Systematic Review on Quality Requirements. Methods Inf Med 2010;49(04):320-36.
http://dx.doi.org/10.3414/ME10-01-0038

25. Pavlenko E, Strech D, Langhof H. Implementation of data access and use procedures in clinical data warehouses. A systematic review of literature and publicly available policies. BMC Med Inform Decis Mak 2020;20(1):157.
http://dx.doi.org/10.1186/s12911-020-01177-z

26. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC*, et al.* Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. Journal of the American Medical Informatics Association : JAMIA 2007;14(1):1-9.
http://dx.doi.org/10.1197/jamia.M2273

27. Commission nationale informatique et des libertés. Référentiel relatif aux traitements de données à caractère personnel mis en oeuvre à des fins de création d'entrepôts de données dans le domaine de la santé. Paris: CNIL; 2021.
https://www.cnil.fr/sites/default/files/atoms/files/referentiel_entrepot.pdf

28. I2B. i2b2 Common Data Model Documentation - Bundles and CDM - i2b2 Community Wiki [En ligne] 2022.
https://community.i2b2.org/wiki/display/BUN/i2b2+Common+Data+Model+Documentation

29. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S*, et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-30.
http://dx.doi.org/10.1136/jamia.2009.000893

30. Schuemie M. The Book of OHDSI. ; 2021.
https://ohdsi.github.io/TheBookOfOhdsi/

31. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ*, et al.* Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-8.
http://dx.doi.org/10.3233/978-1-61499-564-7-574

32. Sentinel Common Data M. Sentinel Common Data Model \textbar Sentinel Initiative [En ligne] 2022.
https://www.sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

33. Psaty BM, Breckenridge AM. Mini-sentinel and regulatory science . Big data rendered fit and functional. N Engl J Med 2014;370(23):2165-7.
http://dx.doi.org/10.1056/NEJMp1401664

34. Pcornet. Data PCORnet. The National Patient-Centered Clinical Research Network [En ligne] 2022.
https://pcornet.org/data/

35. Bourke A, Bate A, Sauer BC, Brown JS, Hall GC. Evidence generation from healthcare databases: recommendations for managing change. Pharmacoepidemiology and Drug Safety 2016;25(7):749-54.
http://dx.doi.org/10.1002/pds.4004

36. Pasco J, Campillo-Gimenez B, Guillon L, Cuggia M. Pré-screening et études de faisabilité : l'apport des entrepôts de données de cliniques. Rev Epidemiol Sante Publique 2019;67:S96.
http://dx.doi.org/10.1016/j.respe.2019.01.068

37. Hernán MA. Methods of Public Health Research — Strengthening Causal Inference from Observational Data. N Engl J Med 2021;385(15):1345-8.
http://dx.doi.org/10.1056/NEJMp2113319

38. Madec J, Bouzill G, Riou C, Van Hille P, Merour C, Artigny M-L*, et al.* eHOP Clinical Data Warehouse: From a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019:1536-7.
http://dx.doi.org/10.3233/SHTI190522

39. Chazard E, Balaye P, Balcaen T, Genin M, Cuggia M, Bouzillé G*, et al.* Book music representation for temporal data, as a part of the feature extraction process: A novel approach to improve the handling of time-dependent data in secondary use of healthcare structured data. Stud Health Technol Inform 2022;290:567-71.
http://dx.doi.org/10.3233/SHTI220141

40. Goldacre B, Morley J, Hamilton N. Better, Broader, Safer: Using Health Data for Research and Analysis : Secretary of State for Health and Social Care; 2022.
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1067058/summary-goldacre-review-using-health-data-for-research-and-analysis.pdf

41. Plateforme des Données de Santé . Referentiel de securite SNDS - Guide d'ccompagnement : HDH; 2020.
https://www.snds.gouv.fr/download/Guide_accompagnement.pdf

42. Paris N, Doutreligne M, Parrot A, Tannier X. Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. TALMED 2019 : Symposium satellite

francophone sur le traitement automatique des langues dans le domaine biomédical. Lyon; 2019.
https://hal.archives-ouvertes.fr/hal-02564721/document

43. European Commission. Study about the impact of open source software and hardware on technological independence, competitiveness and innovation in the EU economy. Brussels: European Commission; 2021.
https://digital-strategy.ec.europa.eu/en/library/study-about-impact-open-source-software-and-hardware-technological-independence-competitiveness-and

44. Direction interministerielle des systemes d'information et de communication, Ayrault JM. Orientations pour l'usage des logiciels libres dans l'administration. Légifrance - Droit national en vigueur - Circulaires et instructions. Paris: DISIC; 2012.
https://www.legifrance.gouv.fr/download/file/pdf/cir_35837/CIRC

45. Nagle F. Government technology policy, social value, and national competitiveness. Rochester, NY: Harvard Business School; 2019.
https://www.hbs.edu/ris/Publication%20Files/19-103_70f212c8-c4fe-4989-ac99-e03cf8bbf02d.pdf

46. Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. Article 16 : Journal officiel 8 octobre 2016; 2016.
https://www.legifrance.gouv.fr/jorf/jo/2016/10/08/0235

47. Circulaire n°6264/SG du 27 avril 2021 relative à la politique publique de la donnée, des algorithmes et des codes sources [En ligne] 2021.
https://www.legifrance.gouv.fr/circulaire/id/45162

48. Shang N, Weng C, Hripcsak G. A conceptual framework for evaluating data suitability for observational studies. Journal of the American Medical Informatics Association: JAMIA 2018;25(3):248-58.
http://dx.doi.org/10.1093/jamia/ocx095

49. Looten V, Kong Win Chang L, Neuraz A, Landau-Loriot M-A, Vedie B, Paul J-L*, et al.* What can millions of laboratory test results tell us about the temporal aspect of data quality? Study of data spanning 17 years in a clinical data warehouse. Comput Methods Programs Biomed 2019;181:104825.
http://dx.doi.org/10.1016/j.cmpb.2018.12.030

50. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD*, et al.* Datasheets for datasets. Communications of the ACM 2021;64(12):86-92.
http://dx.doi.org/10.1145/3458723

51. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys 2021;54(6):115:1-:35.
http://dx.doi.org/10.1145/3457607

52. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B*, et al.* Model Cards for Model Reporting. 2019/01// 2022/06/14/. 2019. 220-9.
http://arxiv.org/abs/1810.03993

# Contributions des auteurs et remerciements

## Groupe de travail

- Matthieu Doutreligne a conçu et mené le projet, analysé la littérature, défini les catégories d'études, catégorisé les études, rédigé le guide d'entretien, mené les entretiens avec les invités et rédigé le rapport
- Adeline Degremont a assisté aux entretiens, défini les catégories d'études, rédigé le guide d'entretien et relu le rapport
- Pierre-Alain Jachiet a conçu le projet, rédigé et relu le rapport
- Xavier Tannier a assisté aux entretiens et relu le rapport
- Antoine Lamer a relu le rapport
- Mireille Cecchin a effectué la recherche bibliographique initiale
- Laure Frigère a vérifié et mis en forme la bibliographie
- Christine Mayol a organisé les rendez-vous avec les experts interviewés

## Relecture

- Judith Fernandez, adjointe à la directrice, DEAI – HAS
- Pierre Liot, chef de projet, HAS
- Bastien Guerry, Etalab, direction interministérielle du numérique
- Aude-Marie Lalanne Berdouticq, chercheuse post-doctorale – ENS, Institut santé numérique en société
- Albane Miron de L'Espinay, adjointe au chef de bureau Innovation et Recherche clinique – DGOS, ministère de la Santé et de la Prévention
- Caroline Aguado, stratégie accélération santé numérique au bureau systèmes d'information des acteurs de l'offre de soins (PF5) – DGOS, ministère de la Santé et de la Prévention

## Remerciements

La HAS tient à remercier l'ensemble des participants cités ci-dessus ainsi que l'ensemble des experts interviewés lors des entretiens.

# Abréviations et acronymes

| | |
|---|---|
| **AP-HM** | Assistance publique-Hôpitaux de Marseille |
| **AP-HP** | Assistance publique-Hôpitaux de Paris |
| **API** | *Application Programming Interface,* interface de programmation d'application : normalisation des informations mises à disposition par un système informatique pour échanger avec un autre |
| **ARS** | Agence régionale de santé |
| **ATIH** | Agence technique de l'information sur l'hospitalisation |
| **CHU** | Centre hospitalo-universitaire |
| **CLCC** | Centre de lutte contre le cancer |
| **DGOS** | Direction générale de l'Offre de soins |
| **DIM** | Direction de l'information médicale |
| **DPI** | Dossier patient informatisé |
| **DRCI** | Direction de la recherche clinique et de l'innovation |
| **DSI** | Direction des systèmes d'information |
| **eCRF** | *electronic Case Report Form* : questionnaire électronique utilisé en recherche clinique pour collecter de la donnée |
| **EDS** | Entrepôt de données de santé |
| **EDSH** | Entrepôt de données de santé hospitaliers |
| **EHR** | *Electronic Health Record* : traduction anglaise de dossier de santé électronique *Medical Record* : traduction anglaise de dossier médical patient |
| **EMR** | *Electronic Medical Record* : traduction anglaise de dossier médical patient |
| **FDA** | *Food and Drug Administration* |
| **HAS** | Haute Autorité de santé |
| **HCL** | Hospices civils de Lyon |
| **PDS** | *Plateforme des Données de Santé* |
| **IQSS** | Indicateur de qualité et de sécurité des soins |
| **NICE** | *National Institute for Health and Care Excellence* |
| **PMSI** | Programme de médicalisation des systèmes d'information |
| **RIHN** | Référentiel des actes innovants hors nomenclatures |
| **SI** | Système d'information |
| **SIH** | Système d'information hospitalier |

Retrouvez tous nos travaux sur
www.has-sante.fr

HAUTE AUTORITÉ DE SANTÉ

Développer la qualité dans le champ sanitaire, social et médico-social