Panorama des entrepôts de données santé hospitaliers (EDSH) dans les CHU/CHR de France

Matthieu Doutreligne[1,2], P.A. Jachiet[1], A. Degremont[1], A. Lamer[3,4], X. Tannier[5]

[1]Mission Data, Haute Autorité de Santé, Saint-Denis -- [2]SoDa team, Inria, Palaiseau -- [3]METRICS, ULR 2694, Univ. Lille, CHU Lille -- [4]F2RSM Psy Hauts-de-France -- [5]Limics, Sorbonne Université, Inserm, Paris

HAUTE AUTORITÉ DE SANTÉ

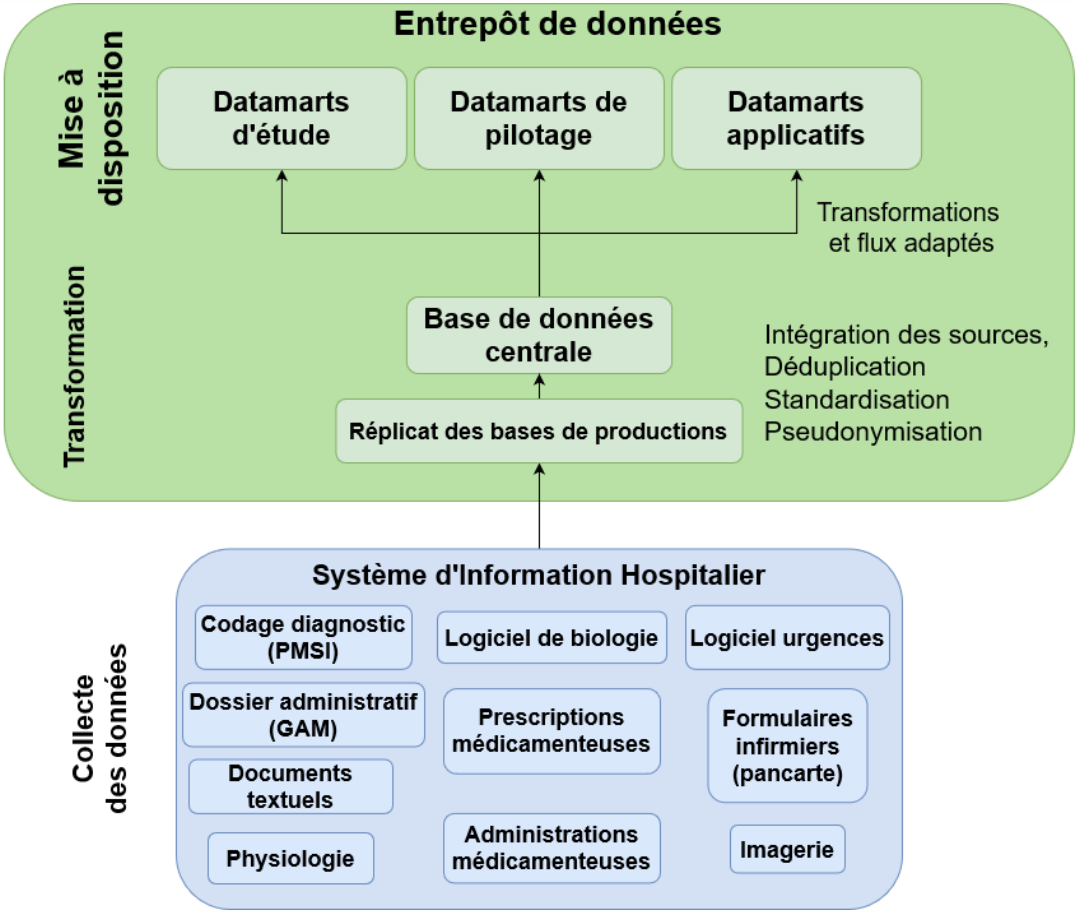
17/03/2023

L’ EDSH : données hospitalières de vie réelle

Mise en commun des données d’un ou plusieurs systèmes d’informations médicaux,

Sous un format homogène

Pour des réutilisations à des fins de pilotage, de recherche ou dans le cadre des soins.



Trois étapes de structuration depuis le Système d’Information Hospitalier (SIH)

- **Collecte,**
- **Transformation**
- **Mise à disposition**

Enquête auprès des EDSH

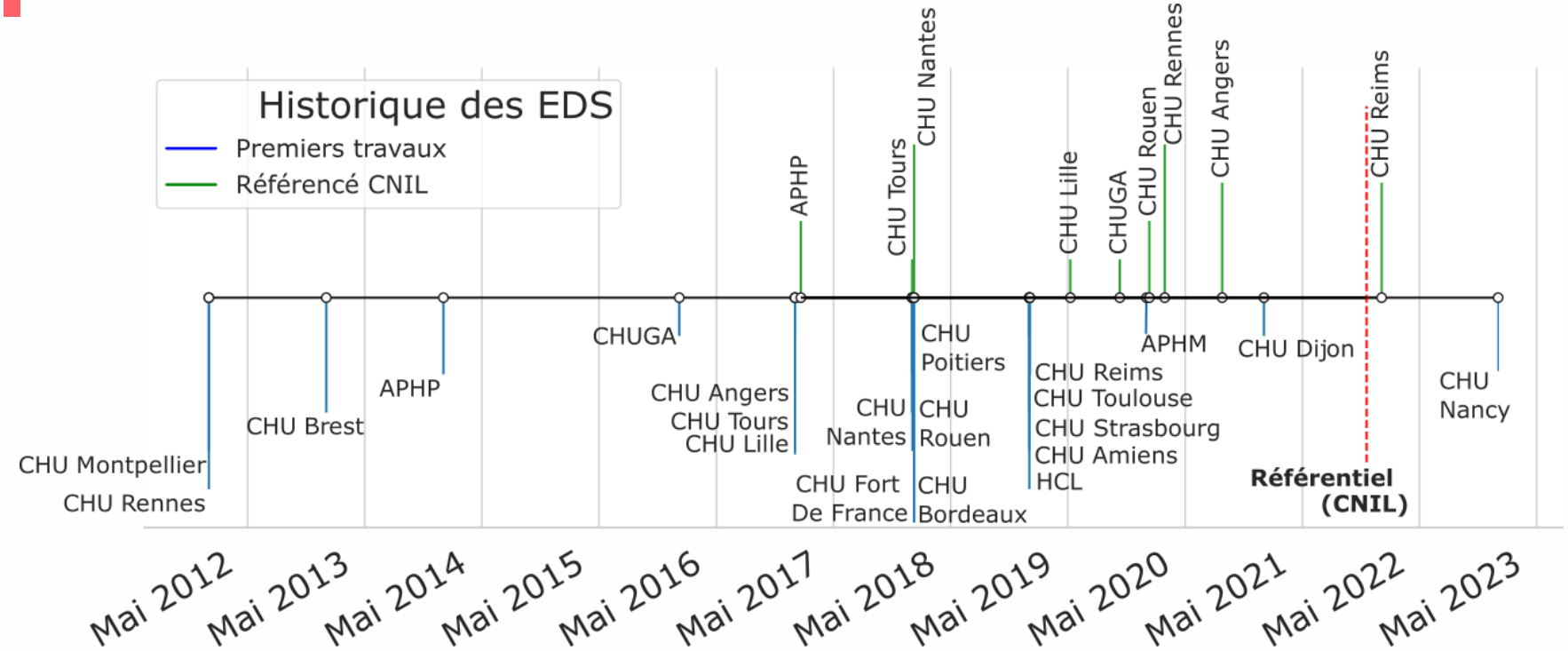
Entretiens semi-structurés avec

- 22 CHU/CHR,
- 4 hôpitaux
- 1 Centre de Lutte Contre le Cancer
- 2 startups
- 1 association
- 3 Institutions

Mail pour les 10 autres CHU/CHR

Dénominateur des résultats:
21 CHU/CHR au moins prospectifs

Temporalité des EDSH



L'EDSH est initié par :
- Un individu moteur (biostats/DIM et appétence info ++)
- Soutenu par la direction ou un clinicien

1. Nécessité d'une équipe EDSH dédiée et pérenne

De taille variable (0,5 à 70 ETP, médiane=7,5) intégrée au sein du DIM, de la DCRI ou d'une équipe de santé publique, créant des **liens transversaux** avec d'autres unités – DSI, cliniciens, DG.

Collecte et harmonisation des données

Etudes de qualité des données
(trop peu valorisées)

Documentation et diffusion des connaissances *(complexe)*

Mise à disposition et accompagnement pour l'export des données

Plateforme d'analyse *(datalab, 6/21)*

Cœur de métier

Développement de **codes d'analyse**

Applicatifs métiers

Outils dédiés au pilotage (administratifs ou chefs de services)

Services complémentaires

2. Vers une gouvernance à 3 niveaux

Local

- Structuration des données, usages et besoins terrains
- Sur le modèle du Centre de Données Cliniques

Interrégional

- Réseaux d'entrepôts : HUGO (6 CHU), AP-HP, Hauts-de-France (), Grand-Est ()
- Mutualisation des solutions techniques et des compétences, groupes de travail thématiques

National

- Coordination, mise à disposition d'outils communs, incitations, doctrines et méthodologies, schémas données, appui juridique

Données

Type de données	Nombre d'EDSH	Ratio
GAM	21	100 %
PMSI	20	95 %
Textes ++	20	95 %
Biologie	20	95 %
Circuit du médicament	16	76 %
Imagerie	4	19 %
Pancarte	4	19 %
Anatomopathologie	3	14 %
Réanimation	2	10 %
Dispositifs médicaux	2	10 %

La complexité d'un **EDSH est le reflet de celle du SIH**

Les données sont correctement renseignées si elles **servent aux soignants**

3.📣 Standards de données et socle commun

Constats

💬 **Modèles communs de données** : Connaissance des principaux modèles (OMOP 7/21), mais pas d'entente global (sauf eHop dans l'Ouest)

🈶 **Nomenclatures communes** : Gros problèmes de sémantique, peu d'outils et pas d'instance nationale pour aider et coordonner

📣 Pistes

🌍 Favoriser le plus possible les **modèles internationaux éprouvés et open source**, développer une coordination nationale sur ce sujet

🌱 Proposer un **socle commun de données** avec des métadonnées

Usages

Etudes et Recherche (Tous)

- **Recherche interventionnelle (études de pré-screening)**
- **Thèses d'internat**
- Projets de recherche épidémiologiques
- Prédictions cliniques / aide à la décision (🛠)

Pilotage (16/21)

- **Optimisation du codage**
- Tableaux de bord d'activité
- Indicateurs de qualité
- Pharmacovigilance

Usages cliniques (DPI amélioré) (13/21)

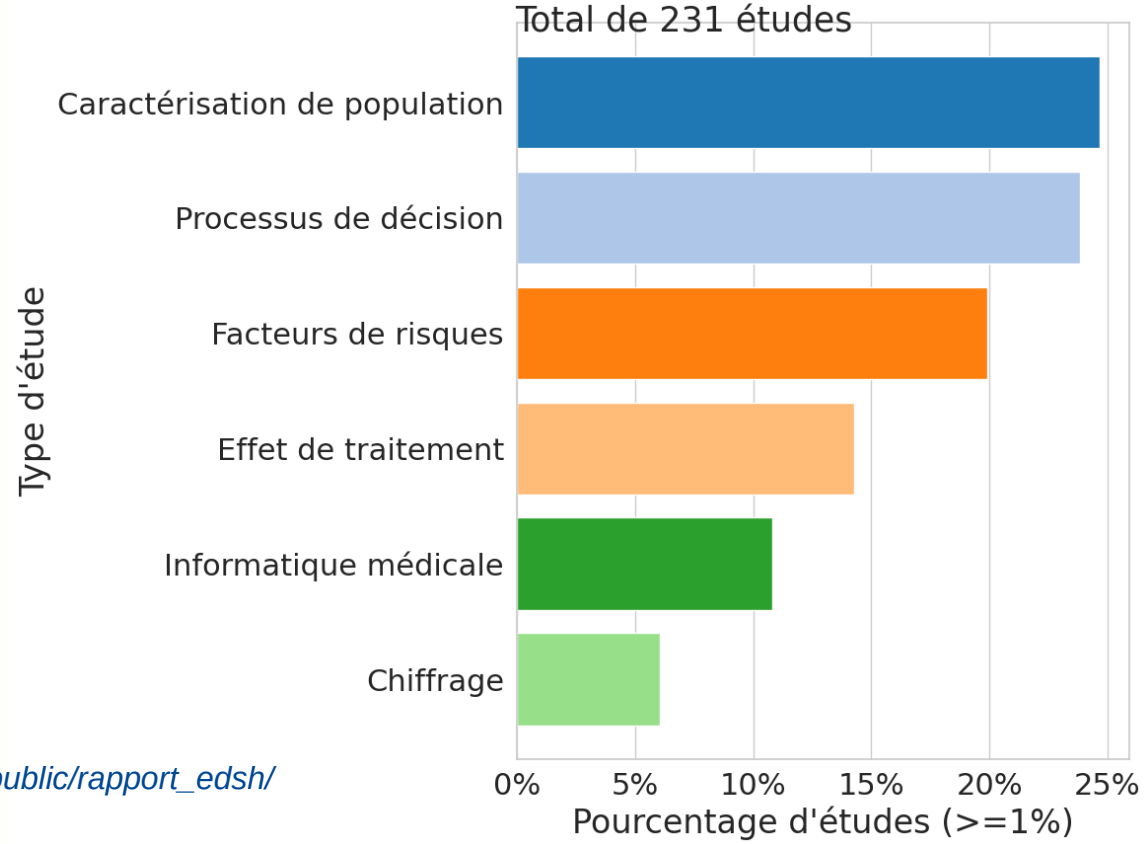
- Automatisation de tâches répétitives (ex: tri auto des prescriptions hospitalières)
- Usages en prévention (ex: inclusion auto dans la filière fracture/ostéoporose)
- Patients similaires pour l'aide au diagnostic (maladies rares)
- Suivi et tri patients pour la coordination (🛠)

Usage recherche : Distribution des études par type

Recensement des études en cours à partir de 9 portails disponibles

Proposition de catégories d'études (inspirées du consortium OHDSI)

Faciliter la compréhension des objectifs d'études et des méthodologies



Données: https://gitlab.has-sante.fr/has-sante/public/rapport_edsh/

4. Pour un écosystème transparent

Pour les patients

Via les portails d'études : à développer et maintenir à jour

(8/14 EDSH avec des études)

Pour l'information scientifique

Nécessité d'unifier les portails de déclaration

(clinicaltrials.gov, portail études HDH, site CHU)

Pour les utilisateurs des données

Open source : coopération, compréhension des traitements de données

⚠ Nécessité de **gouvernance de l'open source** : valorisation, soutien institutionnel

5. Intégrer le parcours de soin extrahospitalier

Un parcours de soin incomplet dans l'EDSH

- **Outcomes, traitements, facteurs de risque**: biais importants en intra-hospitalier uniquement
- **Organisation des soins** : nécessaire complémentarité entre ville et hôpital

Quelles solutions ?

- **Systématisation des appariements SNDS**
- **EDS de villes :** développement et articulation avec les EDSH

Productions

- **Rapport HAS :**
https://has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france

- *Good practices for clinical data warehouse implementation: a case study in France,* https://arxiv.org/pdf/2302.07074.pdf
(en révision chez Plos digital health)

Sommaire

Slides complémentaires

- Objectifs et contextes
- Définitions
- Formulaire d’entretien et acteurs rencontrés
- Détails sur les résultats
- Perspectives HAS
- Acteurs du projet, relecteurs

Entrepôts de Données de Santé Hospitaliers (EDSH)

-

Objectifs, Contexte, Méthodes

Objectifs du cycle d'échange

Principal : Panorama de l'existant, pour identifier les critères de qualité attendus pour des usages HAS

Rapport : Quels caractéristiques des entrepôts de données de santé hospitalier en France ?

Secondaire : Ebaucher des projets réalisables avec certains entrepôts, centré sur des sujets HAS

Projets expérimentaux sur entrepôts

Axe 2 stratégie données HAS : Usage des données de vie réelle

Données de vie réelle (**R**eal **W**orld **D**ata) :
Données collectées en conditions de **pratique courante** (FDA, 2016, HAS 2021, Nice 2022)

Les principales sources

- Bases de remboursement (SNDS)
- **Dossiers Patients Informatiques** (Electronic Health Records)
  - **hospitaliers**
  - de villes : cf. projet P4DP du CNGE, cegedim Health Data (THIN)
- Registres
- Cohortes
- Données générées par les patients

Ecosystème des EDSH

- **Les établissements hospitaliers**
- **Universitaires** : équipe de santé publique, d'informatique médicale, de statistiques, traitement du langage, imagerie, …
- **Editeurs logiciels métiers :** dxcare (dedalus), orbis, cegedim (ville), easily *(HCLs),* etc.
- **Industriels exploitation de données :** eHop, Codoc, Arkhn, Lifen, etc. (panel intéressant dans l'AMI Santé Numérique)
- **DGOS :** pilote l'Appel A Projet EDSH, réunion de lancement le 12 sept. 2022
- **Health Data Hub :** structure et anime l'écosystème, fournit des services juridiques et techniques

Méthode des entretiens

Entretiens semi-directifs

Fiche entretien guidant l'échange

Enregistrement et retranscription pour consultation

Trois tableaux de données structurés

- / Invités
- Entrepôts
- Etudes en cours sur entrepôt

Les acteurs rencontrés

Nombre total de personnes : 60

Equipes (plusieurs affiliations possibles):

- **Direction de la Recherche :** 4 intervenants
- **DIM ou Santé Publique :** 19 intervenants
- **Equipe Entrepôt :** 31 intervenants
- **DSI :** 10 intervenants

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 : facturations (PMSI), autres données administratives, autres actes, interventions et diagnostics structurés, mesures de biologies structurées, 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) ? Quelle(s) aire(s) thérapeutiques ?
	Quels sont les grands types de projets parmi la liste non-exhaustive suivante : création de cohorte, épidémiologie descriptive, épidémiologie analytique (comparative) avec/sans randomisation, pilotage et tableaux de bords, 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 ? 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, code et documentation accessible ? Présence de personnel dédié voire d'une équipe pour vérifier la qualité des données en continu, 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 ? Comment ce document est il tenu à jour vis-à-vis des évolutions constantes de l'entrepôt ? Sous quelle forme ? 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 ? (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), y-a-t-il une documentation spécifique sur leurs performances ? En ce qui concerne le codage manuel (par exemple labelling), 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 confrontés aux définitions conceptuelles ie. métiers et scientifiques ? Une étude des FPR/TPR par rapport à un standard de référence existe-t-il ? 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)) ? 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 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), 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, études post-inscription (actes, accès précoces oncologie). Evaluation des actes : ex. actes de biologie + imagerie réalisés à l'hôpital, tests génétiques en oncologie et maladies rares.
Echange libre	

Usages des données

Usage primaire : soin des patients

Usage secondaire : pas directement pour la prise en charge du patient

Exemples :

- Consultation des antécédents du patient ->
- Transmission des CR d'un service à l'autre ->
- Recherche épidémiologique, études de pré-screening ->
- Tableaux de bords de l'activité hospitalière ->

- Outils d'apprentissage machine déployés en soin courant -> / ?

Etude monocentrique / multicentrique

Etude monocentrique	Etude multicentrique
Caractéristiques	
- Un seul lieu de collecte - Analyste proche de l'équipe soignante - Parfois plan de collecte défini par l'analyste - Peu d'acteurs et d'intermédiaires - Nomenclatures locales	- Multiples lieux de collectes - Analyste pas dans l'équipe soignante - Analyste ignorant du contexte de saisie de l'information - Plusieurs acteurs et intermédiaires - Nomenclatures hétérogènes
Ex. Thèses d'internat	**Ex.** Etudes du réseau OHDSI : https://data.ohdsi.org/OhdsiStudies/

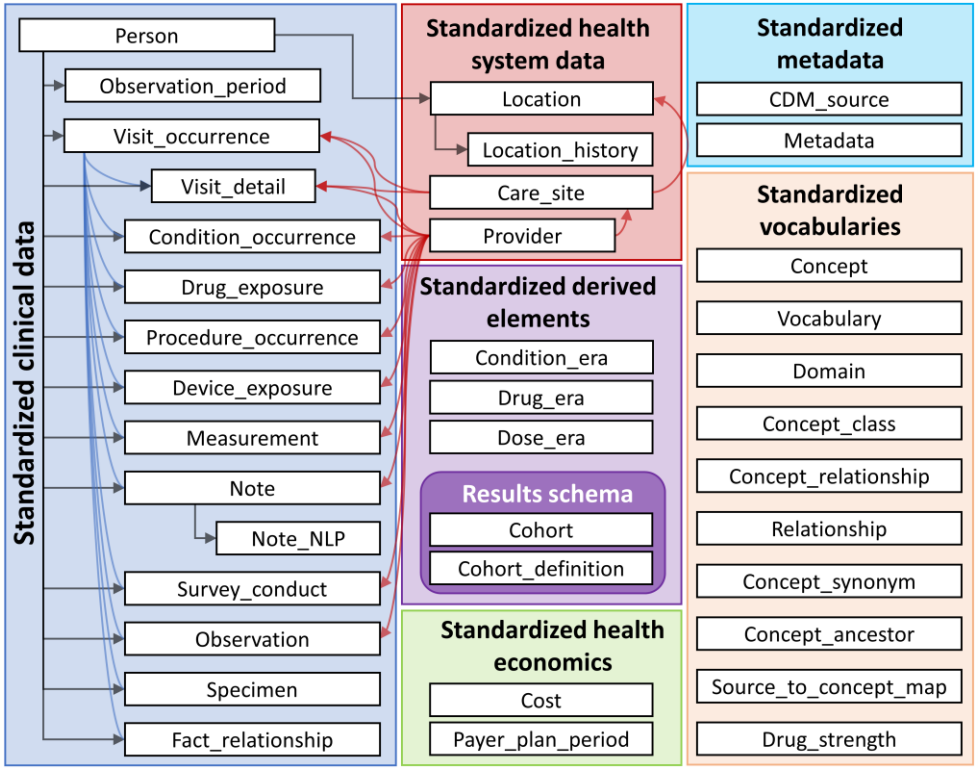
Formats communs de données

OHDSI-OMOP

Standardisation des informations

- **Tables et colonnes**
  https://ohdsi.github.io/CommonDataModel/cdm54.html#Clinical_Data_Tables

- **Nomenclatures**
  https://athena.ohdsi.org/search-terms/terms?query=patient



Formats communs de données

Standardisation des informations

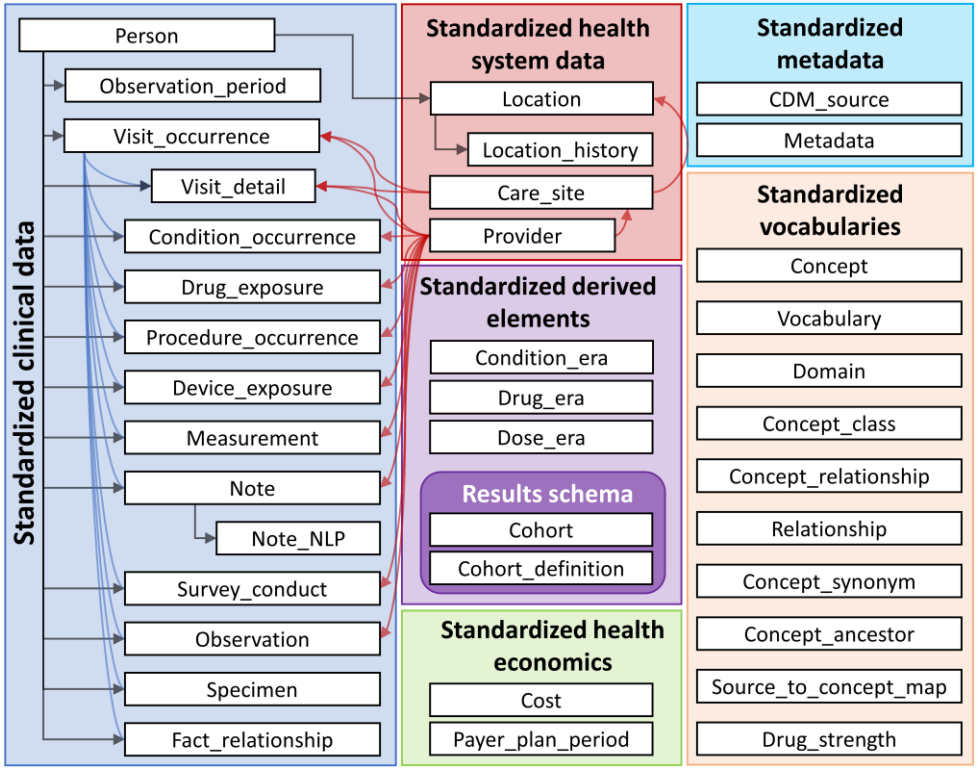
- **Tables et colonnes**
  https://ohdsi.github.io/CommonDataModel/cdm54.html#Clinical_Data_Tables

- **Nomenclatures**
  https://athena.ohdsi.org/search-terms/terms?query=patient

Autres standards en santé

- I2B2, Sentinel, PCORnet
- HL7-FHIR *(transactionnel)*
- UMLS *(ontologie maintenue par le NIH)*

OHDSI-OMOP



Résultats et messages (📣)

-

Gouvernance
Transparence
Données
Usages
Architecture technique
Qualité des données et des méthodes

L’équipe EDSH et son offre de service

De taille variable (0,5 à 70 ETP, médiane=7,5) intégrée au sein du DIM, de la DCRI ou d’une équipe de santé publique, créant des **liens transversaux** avec d’autres unités – DSI, cliniciens, DG.

Collecte et harmonisation des données

Etudes de qualité des données

Documentation et diffusion des connaissances

Mise à disposition et accompagnement pour l’export des données

Plateforme d’analyse *(datalab, 6/21)*

Cœur de métier

Développement de **codes d’analyse**

Applicatifs métiers

Outils dédiés au pilotage (administratifs ou chefs de services)

Services complémentaires

Exigences juridiques complexes : l'exemple de la pseudonymisation

Le triple usage créé des **allers retours juridiques complexes** lors des demandes CNIL

Pseudonymisation

- Pseudonymiser lors de l'ingestion des flux ? -> Perdre l'identité des patients ?
- Dommage lorsque le retour à l'identité est nécessaire: applications de soins et de recrutement, aide au diagnostic, requêteurs transverses de données à destination des cliniciens, outils de prévention.
- Référentiel CNIL (2021): réidentification pour certains cas d'usages où c'est indispensable comme la suppression des données, inclusion dans un essai ou urgences médicales
- Qualité minimale à atteindre pour le système de pseudonymisation ?

Quelles séparations entre les SIH et l'EDSH ?

Dépendance des données avec le parc logiciel du SIH

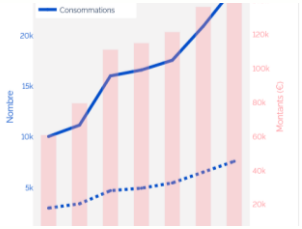
La complexité d'un **EDSH est le reflet de celle du SIH**
-> Ex : Orbis, hétérogénéité des schémas de données pour une même source

La **structuration des schémas de données** (DPI sources) est un problème majeur (APHP, Interhop, CHUGA, Arkhn, Rouen)
-> Pas d'accès aux schémas des éditeurs
-> Spécifications récentes par l'ANS mais très insuffisantes pour l'analyse:
https://esante.gouv.fr/sites/default/files/media_entity/documents/DSR-HOP-DPI-Va1.pdf

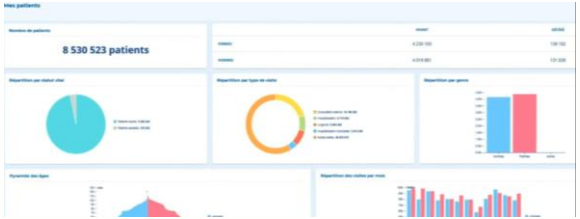
Les données sont correctement renseignées si elles **servent aux soignants**
-> Importance de diriger certaines réutilisations vers les soignants

Types d’études, proposition d’un vocabulaire partagé

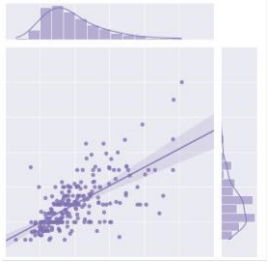
Chiffrage de population



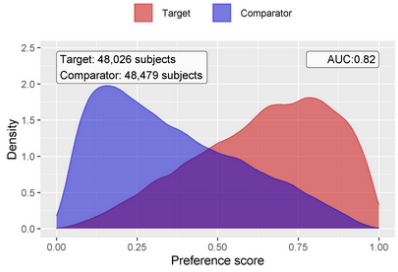
Caractérisation de population



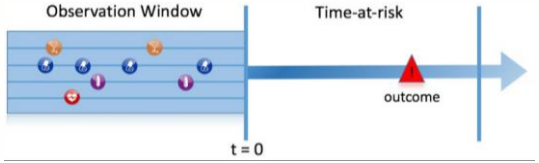
Etude d’association



Effet de traitement

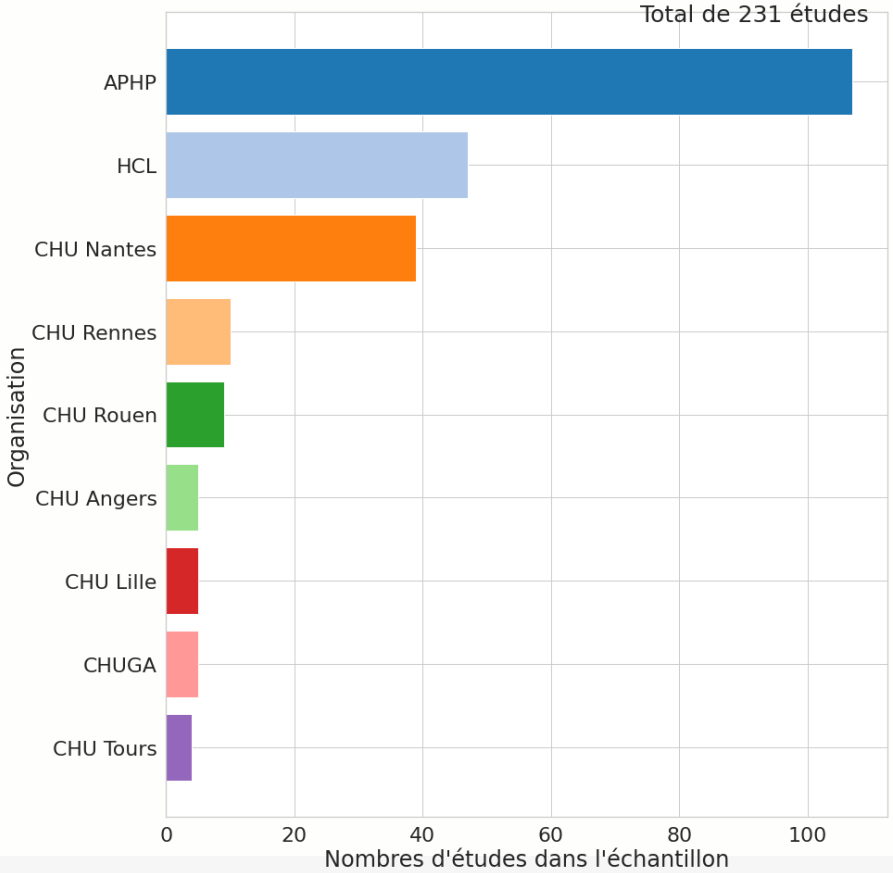


Système d’aide à la décision (ex. pronostic)



Informatique médicale (outils/méthodo)

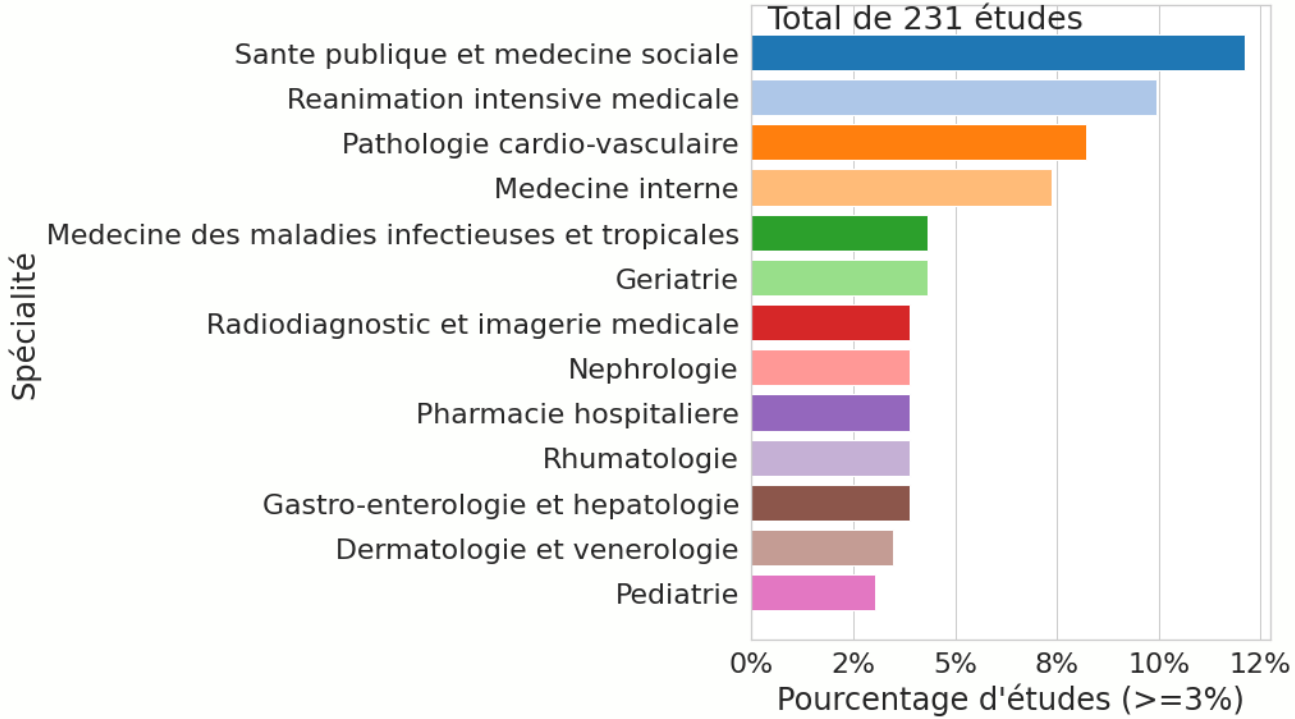
Etudes menées sur entrepôts, un échantillon



Sources:

- Sites internet des entrepôts de données
- Portail d'études en cours pour les EDSH en production: 8/14
- Ajout des HCLs (portail disponible et SIH très homogène).

Distribution des études par spécialité de l'investigateur principal



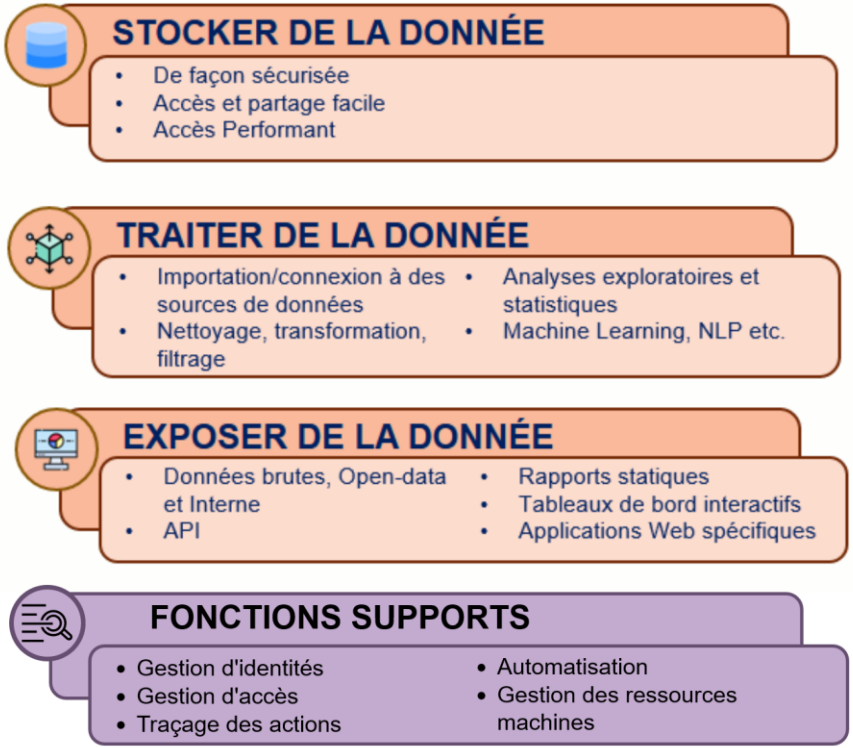
Architecture technique

Beaucoup de **briques technologiques** (parfois >= 35 briques)

Eviter un **éparpillement technologique**
Privilégier 2 /3 chefs de fil nationaux
Publier les schémas d'architecture haut-niveau

Plus de **transparence et de documentation** Favoriser l'open source

Tentative d'externalisation des compétences périlleuse



Architecture technique : Plateforme de données

Fonctions essentielles

Flux de données : connexion et export des sources de données, transformations (nettoyage, aggrégation, filtres, standardisation).

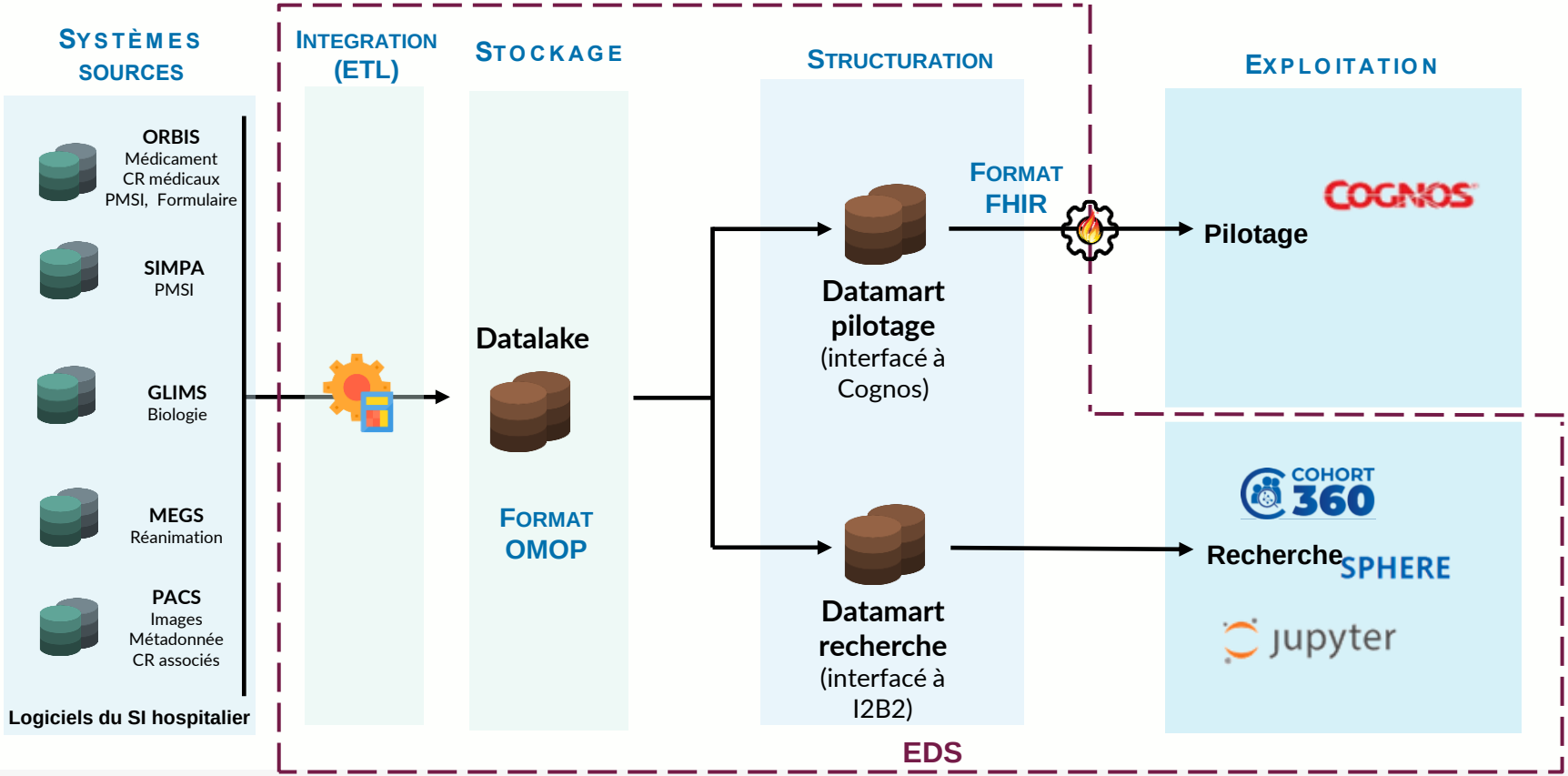
Stockage : moteur de base de données, stockage fichiers, indexations.

Data exposure : Données brutes, APIs, tableaux de bords, environnements de développement, applications webs dédiées.

Fonctions supports

Gestion des identités et des autorisations, traçage, automatization, administration des ressources serveurs.

Exemple de cycle de la donnée, EDS APHP



Principales recommandations

Maillage territorial et lien avec les données de ville

Open source

Equipe dédiée

Gouvernance à 3 niveaux

Vers une meilleure réflexion sur la documentation

De la part des entrepôts, très peu de **documentation en ligne**, sur la donnée, les flux, les procédures d'accès

Valoriser la **documentation ouverte et de qualité**
(lecture à plusieurs niveaux, accessible sur internet, avec des exemples, proche du code, à jour)

Créer des ***dataset cards*** adaptés aux jeux de données des EDSH
Concept utilisé par datagouv.fr et en traitement automatique du langage

Construction de **variables réutilisables :** Inexistant mais besoin identifié pour les études

Perspectives

Expérimentations avec des EDSH partenaires

- Contextualisation de l'**utilisation des actes de génomique en oncologie**
- Développement d'**IQSS basés sur les données EDSH**

Participation au jury pour l'Appel à Projet, Entrepôt de Données de Santé

50 millions sur 3 ans pour le développement des EDS(H)
Fonds: France Relance + Ondam
Partenaires : BPI, DGOS, HDH

Coopération avec une Sociologue de l'Institut Numérique en Santé

Continuités et ruptures dans la « *révolution* » du big data

Aude-Marie Berdouticq

Synthèse et traduction du rapport soumis à Plos digital-health

Activité de biologie hospitalière

Décrire et contextualiser la génomique hospitalière ?

Faisable : La biologie est en grande partie structurée

Temps d'implémentation cours (1 an au maximum, voire plus rapide)

⚠ **Vocabulaires hétérogènes** -> Aide nécessaire au mapping LOINC

⚠ **Génomique non intégrée** -> Mais les codes des gènes testés existent dans le SIH

Statistiques agrégées pour plusieurs établissements :

- Projet fédérateur et porteur pour d'autres usages plus complexes
- Projet d'intérêt pour tout le RIHN (DGOS intéressée)

☑ Aide à la remontée de certains indicateurs qualité

Question : Créer des IQSS à partir des entrepôts ?

Les entrepôts pour le codage automatique (DIM)

- Utilisés souvent pour la valorisation (optimisation codage PMSI)
- Extension en cours à des remontées supplémentaires (article 51 via ATIH)

Les entrepôts pour les indicateurs de qualité (entrepôts, directions cliniques)

- Indicateurs de pilotage (mais très gestion, voire initiatives séparées de l'EDS)
- Indicateurs cliniques : escarres, infections nosocomiales, évaluation de la douleur, hospitalisations non programmées, dénutrition (sarcopénie via imagerie), .

☑ Aide à la remontée de certains indicateurs qualité

Question : Créer des IQSS à partir des entrepôts ?

⏳ Temps d'implémentation long (1 an au minimum)

😨 Choix des sujets (certains impossibles, d'autres faisables)

🤔 Questionnement sur la transposabilité

🤖 L'automatisation des indicateurs est un leurre

🔧 Possibilité d'outils aidant les équipes, vrai potentiel d'amélioration des soins et de gain de temps pour la facturation

⚠ **Proximité + retour équipes cliniques + expériences utilisateurs**

Source ***: Duclos et al., 2020, Effect of monitoring surgical outcomes using control charts to reduce major adverse events in patients: cluster randomised trial, BMJ***

Les acteurs du projet

Groupe de travail

- Pierre-Alain Jachiet, chef ,mission data
- Adeline Degremont, épidémiologiste, mission data
- Xavier Tannier, chercheur en Traitement du Langage
- Antoine Lamer, datascientist
- Matthieu Doutreligne, chef de projet, mission data

Relecteurs

- Judith Fernandez, Adjointe à la directrice, DEAI – HAS
- Pierre Liot, chef de projet, HAS
- Bastien Guerry, Etalab, direction interministérielle du numérique
- 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
- Aude-Marie Lalanne Berdouticq, Chercheuse post-doctorale - ENS, Institut Santé numérique en Société
- *Marc Cuggia, PUPH Informatique Médicale Biostatistique – Université de Rennes*
- *Carole Dorphin, directrice des partenariats au Health Data Hub*
- *Alexis Hecht, chef de projet au Health Data Hub*