

# Les entrepôts de données de santé en France et leur structuration

**Matthieu Doutreligne**

*Haute autorité de Santé, Mission Data*

*Travail effectué dans le cadre d'une thèse à l'Inria, équipe SODA*



**Module de formation sur les Données de Santé**

*12 mars 2025*

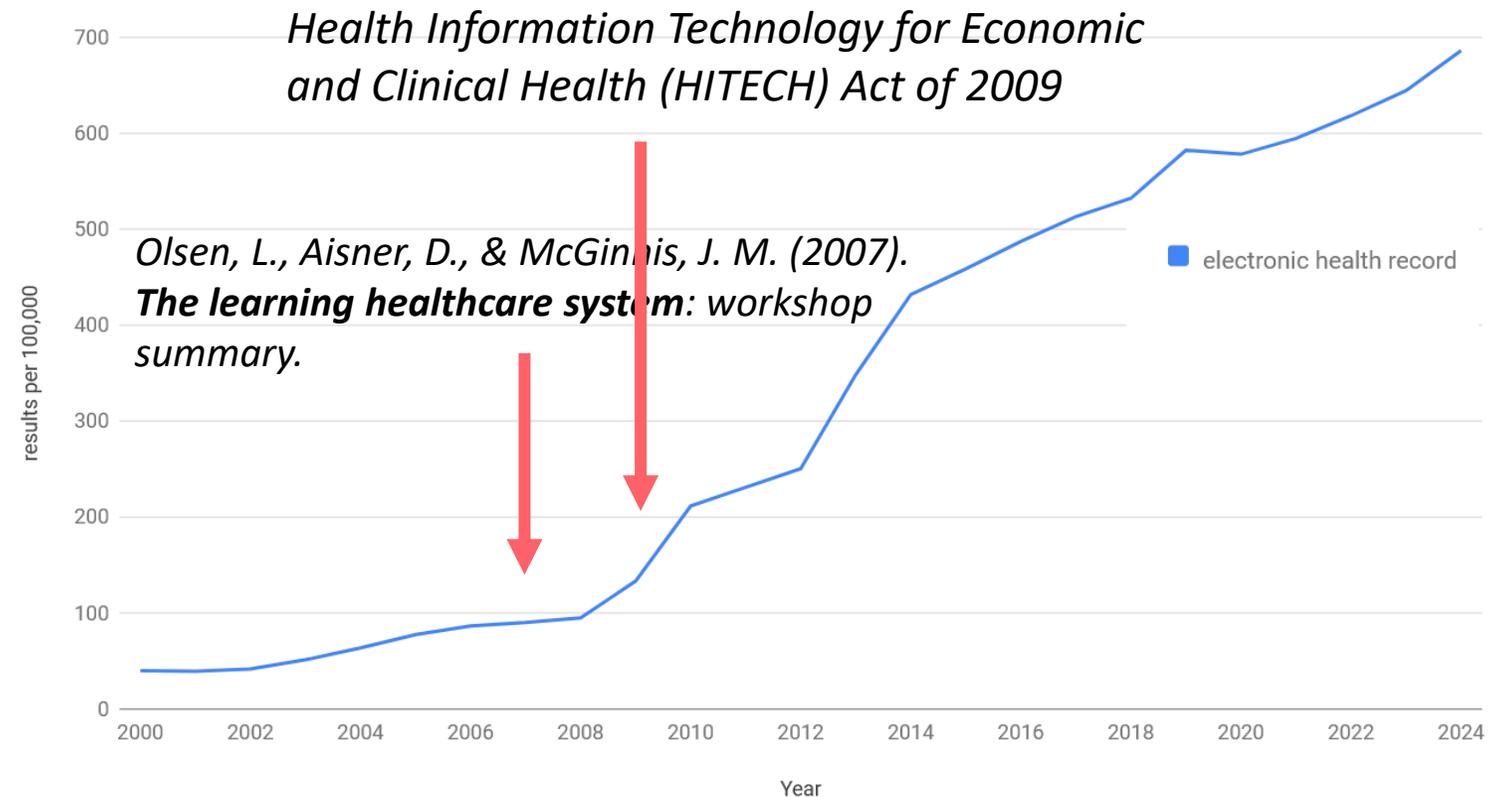


# Dossiers Patients Informatisés : Une adoption croissante

USA :  
De **9,1%** des hôpitaux en **2008** à  
**98.3%** ont un DPI en **2018**

*N. C. Apathy, A. J. Holmgren; J. Adler-Milstein (2021): "A decade post-HITECH: Critical access hospitals have electronic health records but struggle to keep up with other advanced functions". In: Journal of the American Medical Informatics Association*

Results per 100,000 citations in PubMed  
proportion for each search by year, 2000 to 2024



# Axe stratégique données à la HAS : Usage des données de vie réelle

 **Données de vie réelle** (Real World Data) :

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

# Caractéristiques des données de vie réelle

## Avantages pour la santé publique

- Soins courants
- Bonne couverture de la population
- Collection peu couteuse
- Données de grande dimension

## Inconvénients

- Biais de confusion  
(interventions non aléatoires)
- Complexes
- Qualité hétérogène

# Cycle d'échange avec les porteurs des EDSH

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

 **Rapport (nov. 2022) : Quels caractéristiques des entrepôts de données de santé hospitalier en France ?**



## I. Panorama EDSH HAS 2022

1. Gouvernance
2. Données
3. Usages
4. Transparence
5. Plateforme de données

## II. Panorama CNIL 2024

## III. Projet indicateurs qualités HAS

1. Motivation et organisation
2. Méthodes
3. Enseignements
4. Zoom sur l'expérimentation AP-HP

## IV. Conclusions



# I. Panorama EDSH HAS 2022

# L'EDSH : données hospitalières de vie réelle

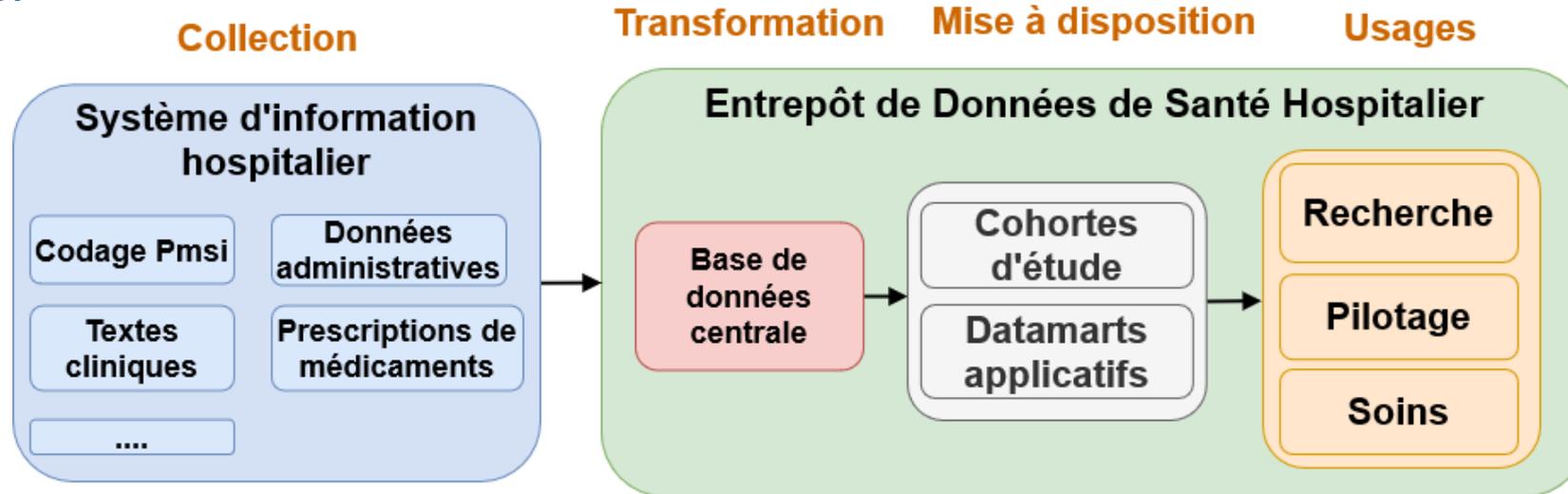
**Mise en commun** des données d'un ou plusieurs systèmes d'informations médicaux,

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



# Etudes monocentriques / multicentriques

## Monocentrique

## 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

**Ex.** Thèses d'internat

- 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.** Etudes du réseau OHDSI :

<https://data.ohdsi.org/OhdsiStudies/>

# Ecosystème des EDSH

- **Les établissements hospitaliers** : professionnels de santé, patients, administratifs
- **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** : Enovacom (eHop), Codoc, Arkhn, Lifen, etc. (panel intéressant dans [l'AMI Santé Numérique](#) )
- **Ministère (DGOS)** : pilote [l'Appel A Projet EDSH](#) (sept. 2022)
- **Health Data Hub** : structure et anime l'écosystème, fournit des services juridiques et techniques

# Méthode pour faire ce panorama

-  Entretiens semi-directifs (mars – novembre 2022)
-  Fiche entretien guidant l'échange
-  Enregistrement et retranscription pour consultation

## Trois tableaux de données structurés

-  /  Invités
-  Entrepôts
-  Etudes en cours sur entrepôt

[https://gitlab.has-sante.fr/has-sante/public/etudes/rapport\\_edsh/-/tree/master/data](https://gitlab.has-sante.fr/has-sante/public/etudes/rapport_edsh/-/tree/master/data)



# 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

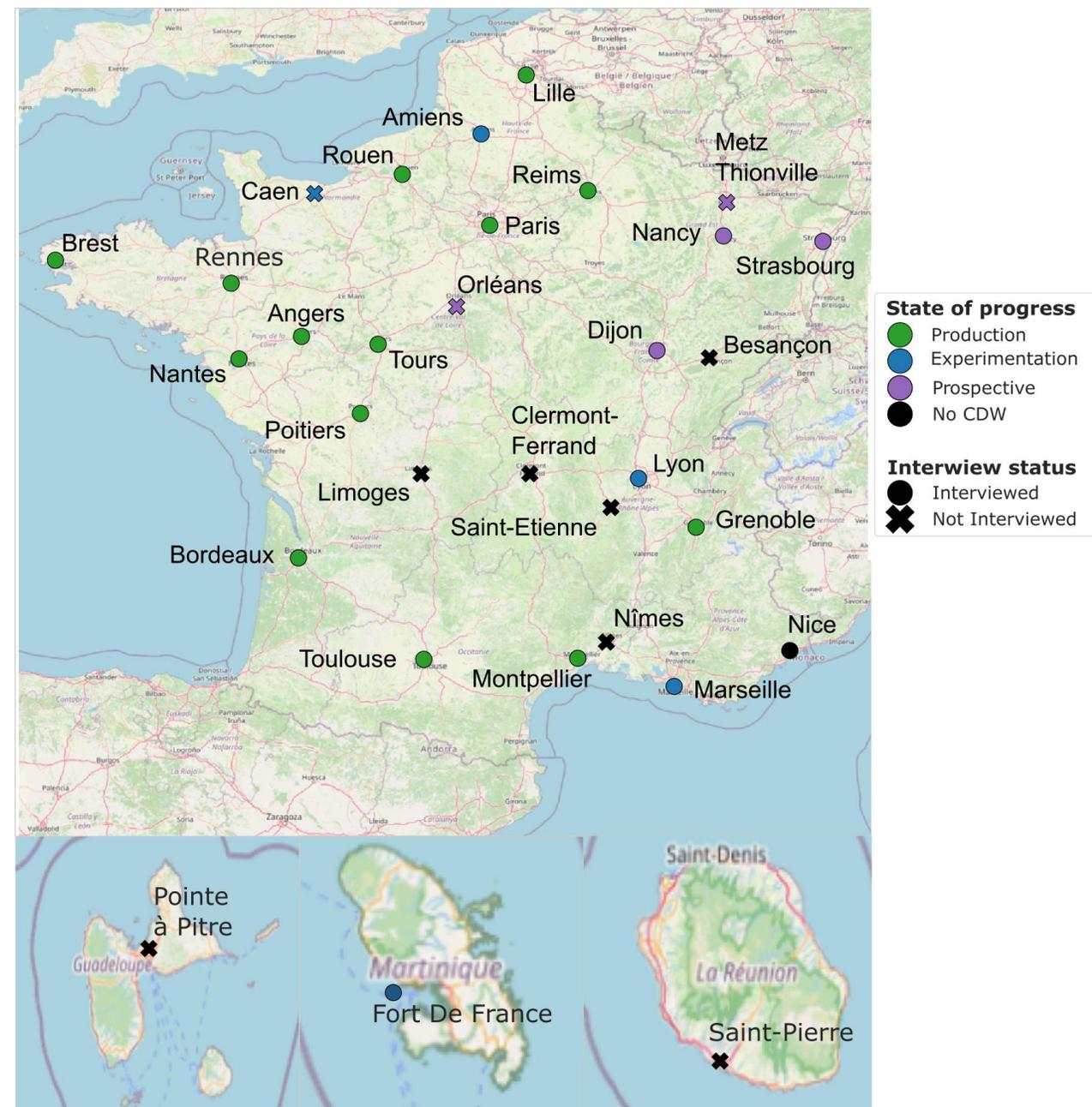
# Enquête auprès des EDSH

Entretiens semi-structurés avec

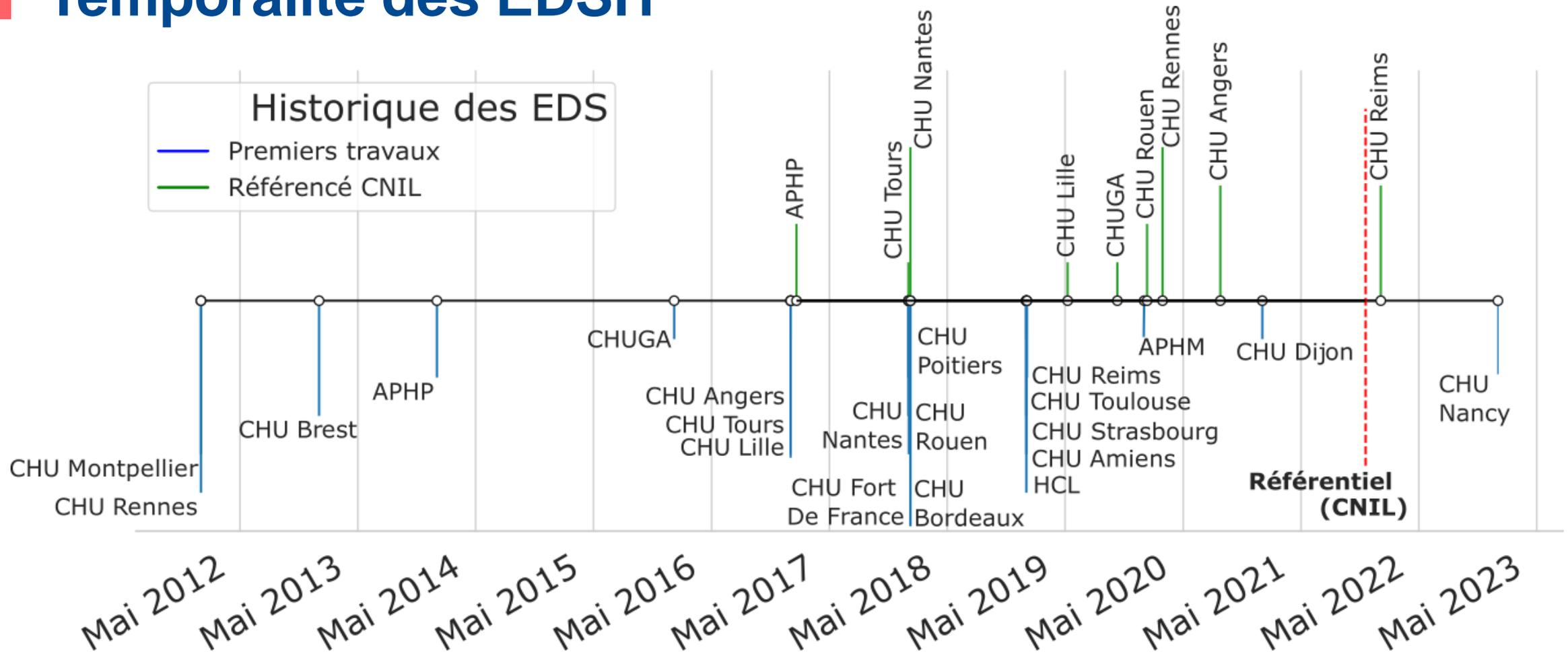
- 22 CHU/CHR,
- 4 hôpitaux
- 1 Centre de Lutte Contre le Cancer
- 2 startups
- 1 association
- 3 Institutions

Mail pour les 10 autres CHU/CHR

Dénominateur des résultats  
21 CHU/CHR au moins prospectifs (interviewés)



# Temporalité des EDSH



L'EDSH est initié par :

-  Un.e individu.e moteur (biostats/DIM et appétence info ++)
-  Soutenu.e par la direction ou un clinicien

# Une équipe EDSH aux multiples fonctions

**Taille variable** (0,5 à 70 ETP, médiane=7,5) intégrée au DIM, à la DCRI ou à 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**  
*(rarement valorisées)*

 **Documentation et diffusion des connaissances**  
*(complexe)*

 **Mise à disposition et accompagnement pour l'export des données**

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

 Développement de **codes d'analyse**

 **Applicatifs métiers**

 **Outils dédiés au pilotage**  
(administratifs ou chefs de services)

**Services complémentaires**

**Cœur de métier**

# Vers une gouvernance à 3 niveaux



## Local

- **Structuration des données**, usages et besoins terrains
- Sur le modèle du [Centre de Données Cliniques](#) (modèles Rennais)



## 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**, appui financier et juridique, de doctrines outils et méthodologies, schémas de données

# 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**

# Standards de données et socle commun

## Standardisation des informations

- **Tables et colonnes**

[https://ohdsi.github.io/CommonDataModel/cdm54.html#Clinical\\_Data\\_Tables](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*)

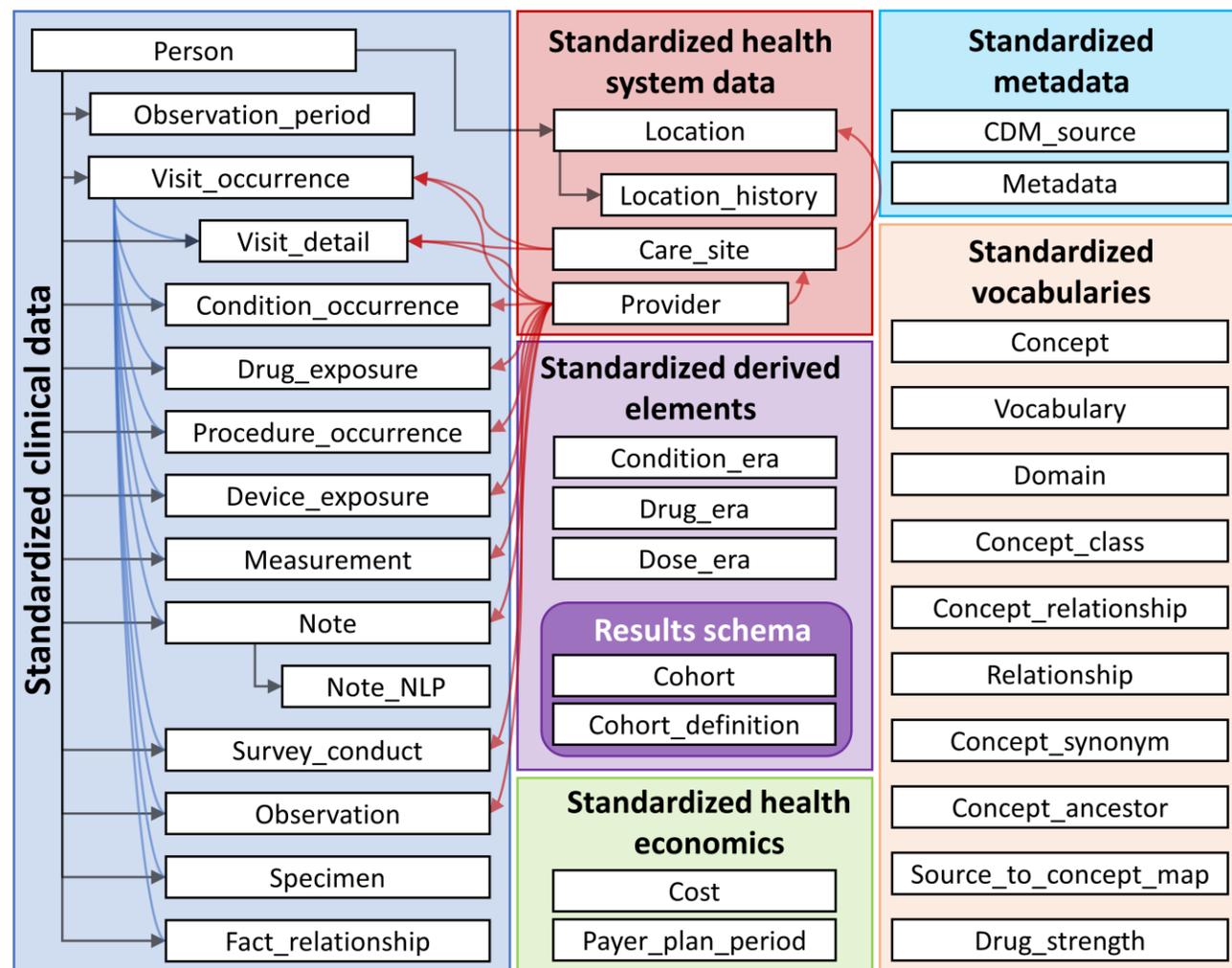


Schéma de **OHDSI-OMOP**

# 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 des données

## Définitions



**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 -> / ?

# Usages des EDSH



## Etudes et Recherche (Tous)

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



## Pilotage (16/21)

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

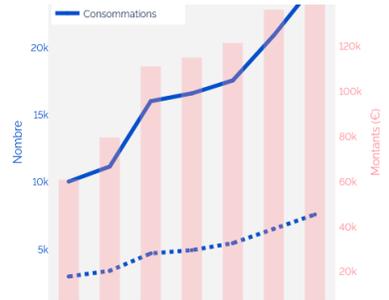


## Usages cliniques (DPI amélioré) (13/21)

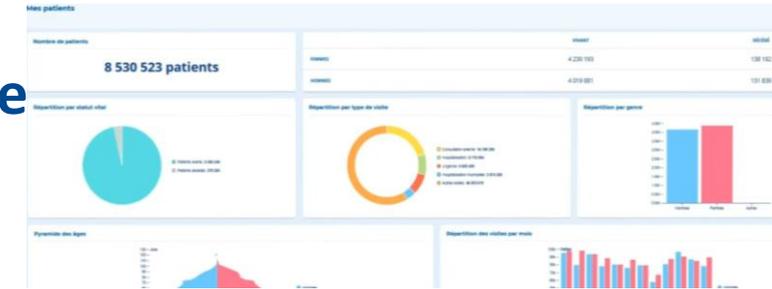
- Automatisation de tâches répétitives (ex: tri auto des prescriptions hospitalières)
- Usages en prévention (ex: inclusion auto dans la filière fracture/ostéoporose)
- Patients similaires pour l'aide au diagnostic (maladies rares)
- Suivi et tri des patients pour la coordination (🔧)

# Types d'études, proposition d'un vocabulaire partagé

Chiffrage de population



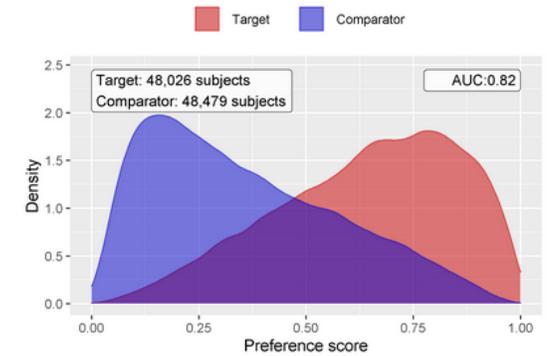
Caractérisation de population



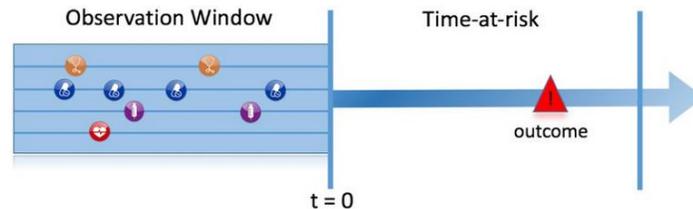
Etude d'association



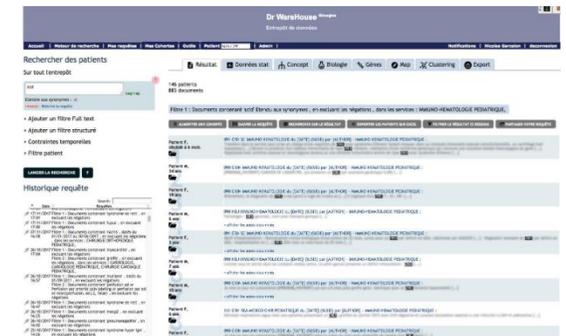
Effet de traitement



Processus de décision (ex. pronostic)



Informatique médicale (outils/méthodo)

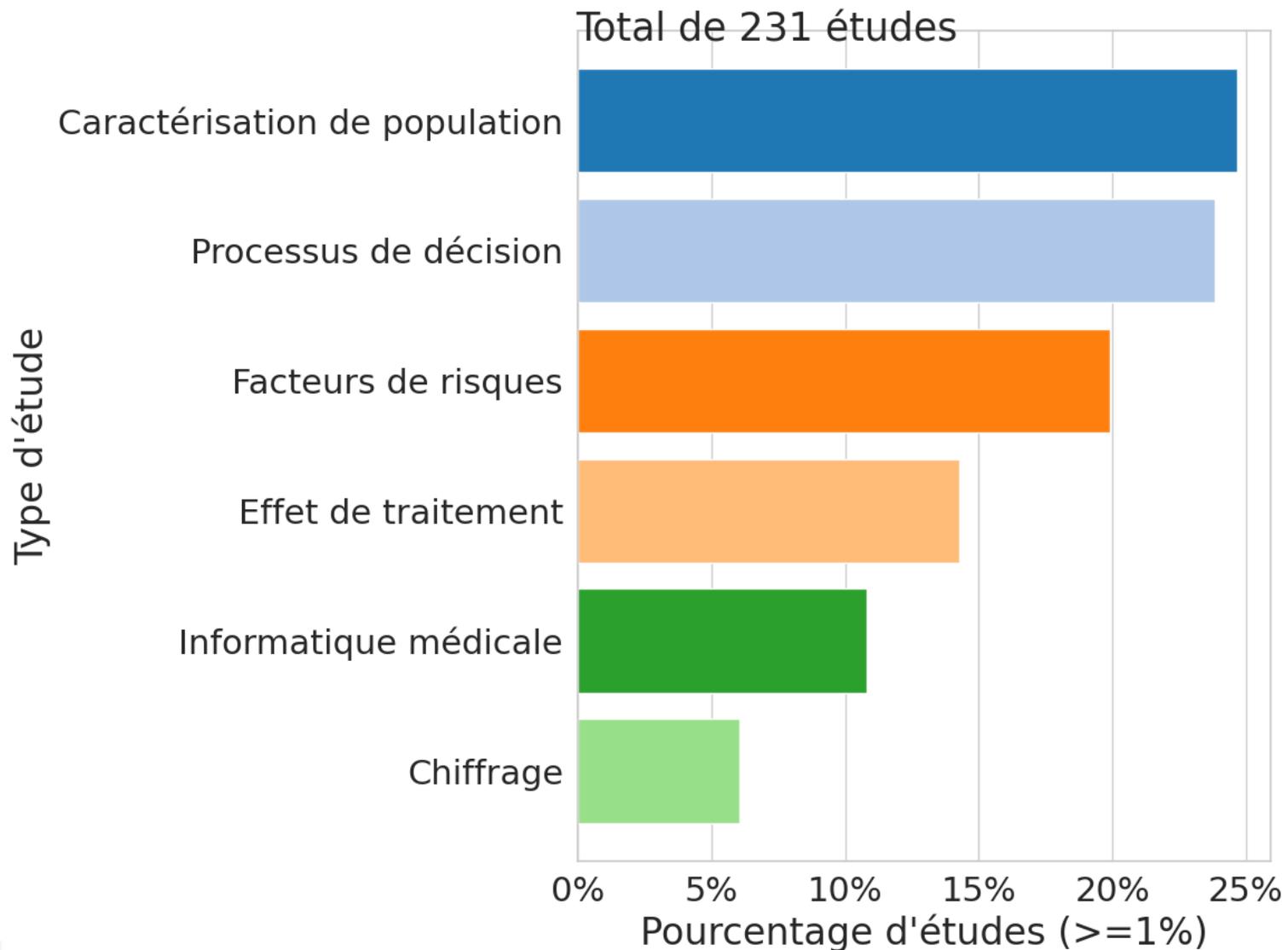


# Usage recherche : Distribution des études par type

 **Recensement des études**  
en cours à partir de 8 portails  
disponibles

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

 **Faciliter la compréhension des**  
objectifs d'études et des  
méthodologies



# Transparence encore insuffisante

## 🙄 Information patient via les portails d'études (référentiel CNIL)

Pour 8/14 EDSH en production, dont 3 manifestement peu à jour

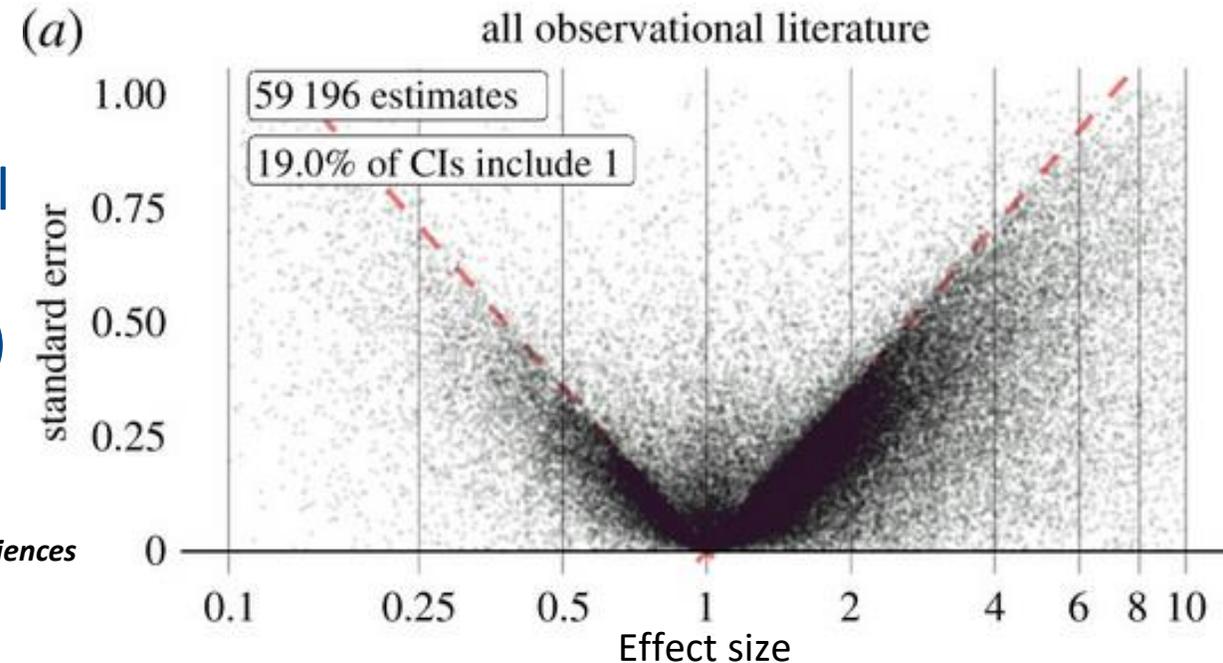
-> **Problème éthique** (lié à une difficulté de mise en œuvre)

## 📖 Information scientifique

Déclaration non unifiée (clinicaltrials.gov, portail études HDH, site CHU)

-> **Problème de p-hacking** (biais de publication)

Réf : Schuemie, M.J., Ryan, P.B., Hripcsak, G., Madigan, D., Suchard, M.A., 2018. Improving reproducibility by using high-throughput observational studies with empirical calibration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 20170356. <https://doi.org/10.1098/rsta.2017.0356>



# Architecture technique : Plateforme de données

## Fonctions essentielles

- **Flux de données** : connexion et export des sources de données, transformations (nettoyage, agrégation, filtres, standardisation).
- **Stockage** : moteur de base de données, stockage fichiers, indexations.

**Exposition des données** : Données brutes, APIs, tableaux de bords, environnements de développement, applications webs dédiées.

## Fonctions supports

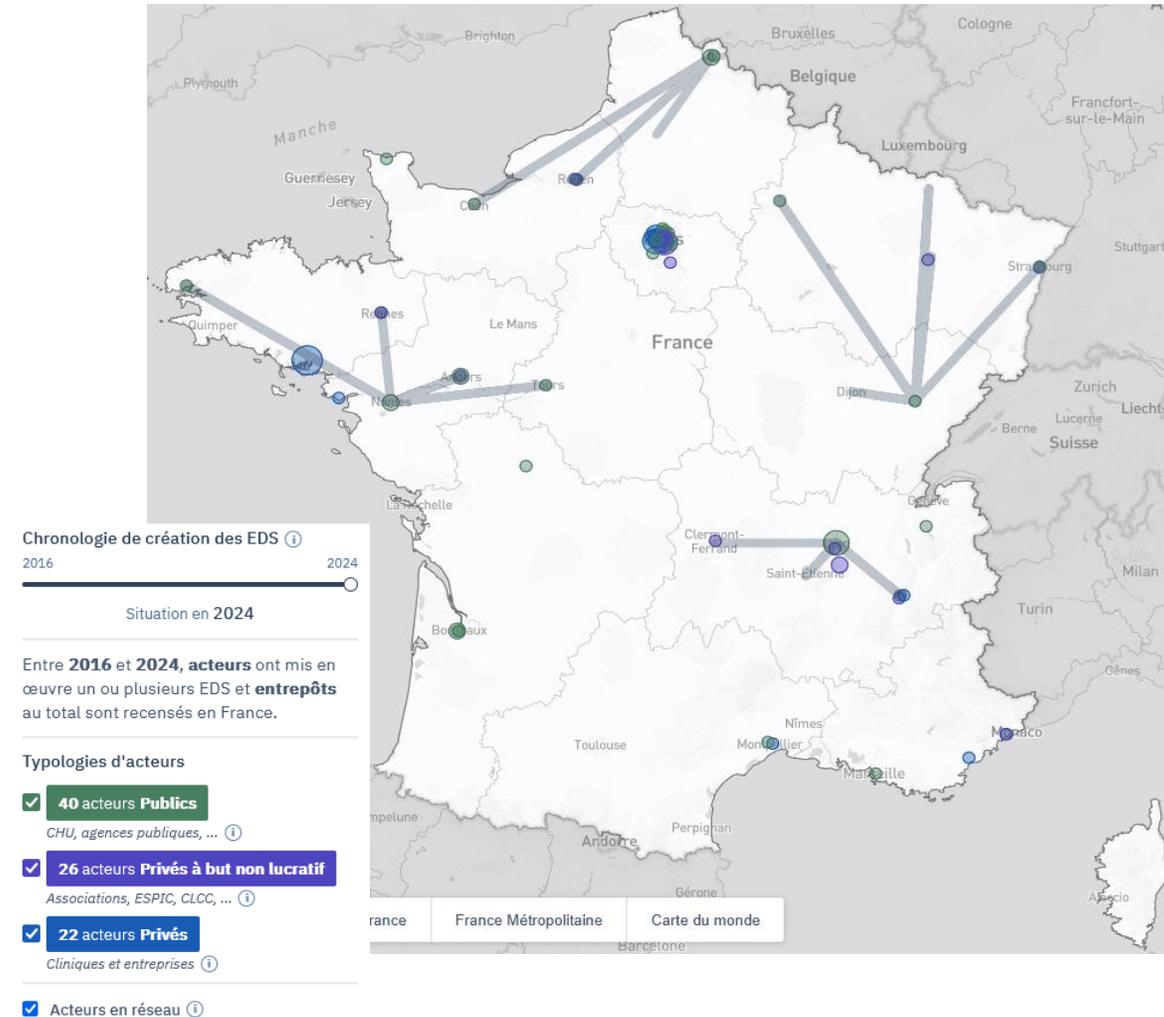
- Gestion des identités et des autorisations, traçage, automatisation, administration des ressources serveurs, capacités de calcul.



## II. Panorama CNIL 2024

# Cartographie 2024 de la CNIL

- Lien web : <https://carto-eds.beta.cnil.fr/>
- Cartographie à partir des **déclarations de conformité CNIL**
- **Mise à jour de notre panorama**
- **Spectre plus large : publics, privés...**
- **Possiblement plusieurs entrepôts par organisme**



# Caractéristiques des EDSH selon le panorama CNIL

## Répartition des acteurs responsables d'entrepôts de données de santé selon le secteur

Publics	<b>40</b>
Privés à but non lucratif	<b>26</b>
Privé	<b>22</b>

## Répartition des responsables d'entrepôts de données de santé selon le type d'acteurs

Centre hospitalier	<b>28</b>
Agence publique	<b>5</b>
Etablissement de santé privé d'intérêt collectif (ESPIC)	<b>3</b>
Enseignement - Recherche	<b>7</b>
Entreprise	<b>20</b>
Association	<b>17</b>
Centre de lutte contre le cancer (CLCC)	<b>6</b>
Clinique	<b>2</b>

## Liste des acteurs ayant plus d'un entrepôt de données de santé (les projets autorisés mais non mis en œuvre sont comptabilisés)

### 5 EDS

- [OpenHealth](#) (dont 3 inactifs)

### 4 EDS

- [HCL](#) (dont 1 inactif)  
- [Unicancer](#)  
- [Clinityx](#)

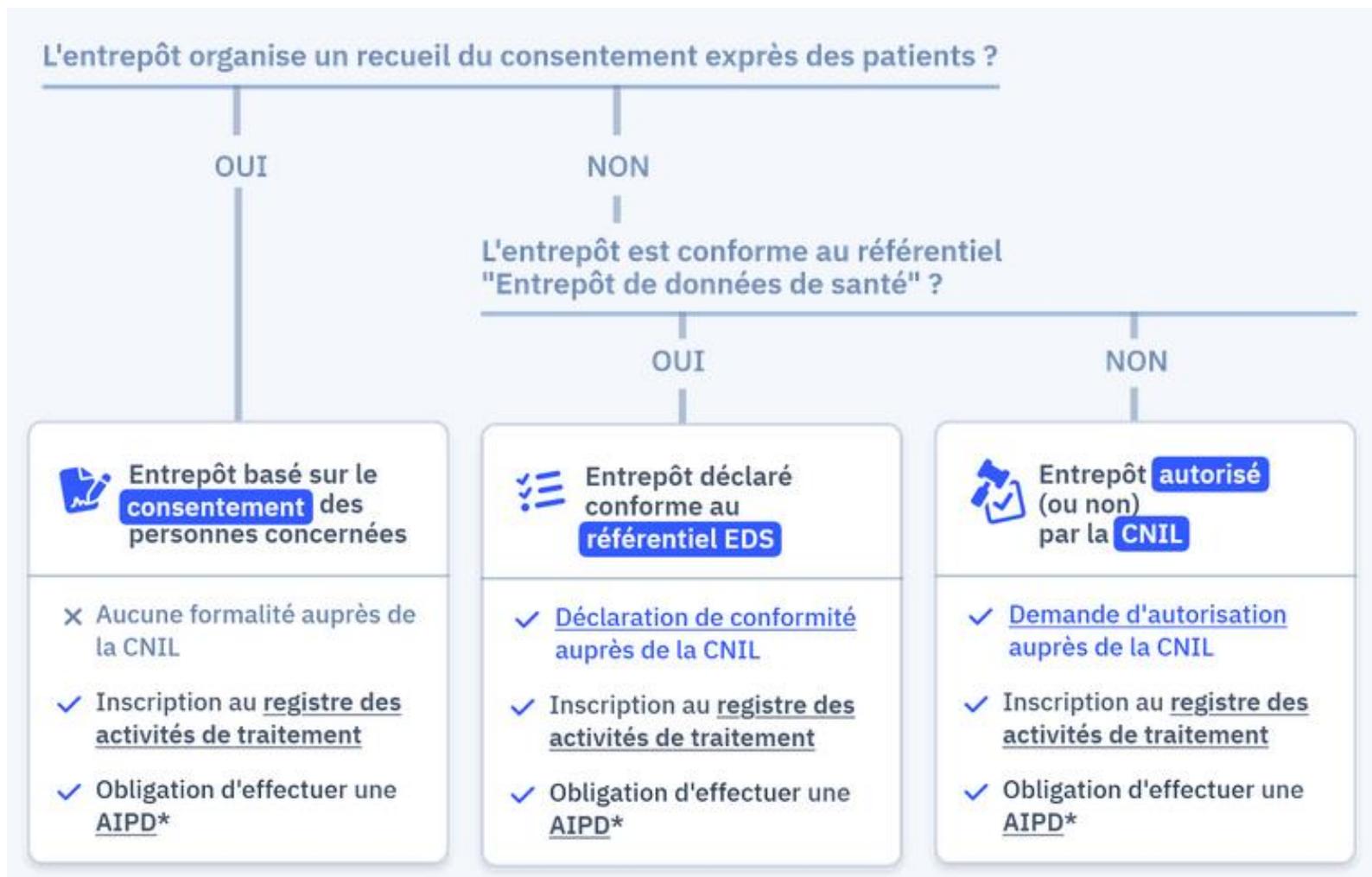
### 3 EDS

- [IQVIA](#) (dont 1 inactif)  
- [INSERM](#)

### 2 EDS

- [AP-HP](#)  
- [CHU Nantes](#)  
- [CHRU Lille](#)  
- [CHU Angers](#)  
- [CHU Bordeaux](#)  
- [Gustave Roussy](#)  
- [RESCUe - RESUVAL](#)  
- [GERS](#)

# Réglementation pour le montage d'un EDSH





### III. **Projet indicateurs qualités HAS**

# Equipes travaillant sur le projet



## CHU Lille

Charlotte Geay  
Dr. Didier Theis  
Dr. Claire Lauérière  
Dr. Grégoire Ficheur



## AP-HP

Olivier Hassanaly  
Pr. Eric Jouvent  
Dr. Penelope Troude



## CHU Bordeaux

Dr. Florence Saillour  
Dr. Vianney Jouhet  
Dr. Francois Rouanet  
Loic Ronflette

## Haute Autorité de Santé

Matthieu Doutreligne  
Agnès Solomiac  
Sandrine Morin  
Pierre-Alain Jachiet  
Dr. Laetitia May-Michelangeli

# Contexte et cas d'usages étudiés

Aujourd'hui,

- une partie des IQSS mesurés à partir des dossiers patients (70 par établissement / an)
- charge de travail importante

Questions à l'origine du projet

Peut-on calculer de **manière automatique** un IQSS sur les données des **EDSH**, avec la même **exigence de qualité** de la mesure ?

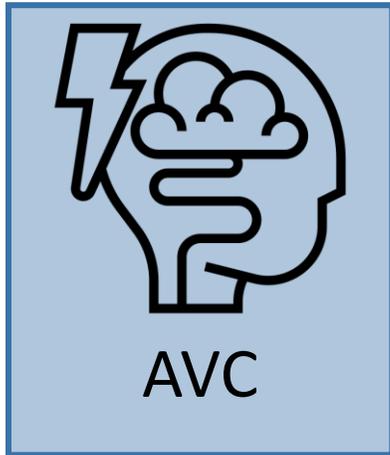
Objectif plus large

Apprendre, ce qui est possible ou non à partir des données des entrepôts

Deux cas retenus

- **Prise en charge de la douleur**
  - Capitalisation sur les travaux existants
- **Délais** dans la prise en charge de **l'AVC**
  - Développement d'un calcul automatique d'un délai

# Thématiques d'intérêt retenue, l'AVC



- **Délai entre l'arrivée dans l'établissement et la 1ère imagerie cérébrale à visée diagnostique**
- Délai entre la date d'arrivée dans l'établissement et la date de l'expertise neuro-vasculaire ou neurologique ou neurochirurgicale
- Délai entre l'arrivée dans l'établissement et la prise en charge rééducative

# Quelques étapes de la méthode



Mai 2023 - Réunions mensuelles pluridisciplinaires 3 EDSH + HAS – déc. 2024



## Calage méthodologique

Calage de choix de sélection des populations, des sources  
Des consignes



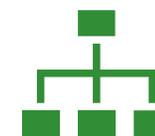
## Partage d'expertise

Discussion de la qualité des sources retenues



## Analyse

Population  
Effectifs obtenus  
Délais observés



## Enseignements généraux

MESURER  
& AMÉLIORER LA QUALITÉ

RAPPORT

Étude de l'intérêt et de la faisabilité de la mesure d'indicateurs de qualité et de sécurité des soins à partir d'entrepôts de données de santé hospitaliers  
Rapport général de faisabilité

Validé par le Collège le 18 décembre 2024

# Exploitation des données : exemple du choix des sources

Quatre sources explorées par Bordeaux pour identifier la population cible

- Diagnostics CIM-10 du PMSI (CIM)
- Comptes-rendus d'imagerie (CR)
- Passage en unité neurovasculaire (UNV)
- Questionnaires de spécialités (V4)

En définitive la CIM-10 est la meilleure source de données

- **Bonne sensibilité** : 99.1 % (1240/1251) des séjours AVC avérés sont repérés par la CIM-10
- **Bonne spécificité** : Seuls 0.3% (4/1244) des séjours AVC repérés par la CIM-10 ne sont pas des séjours AVC avérés

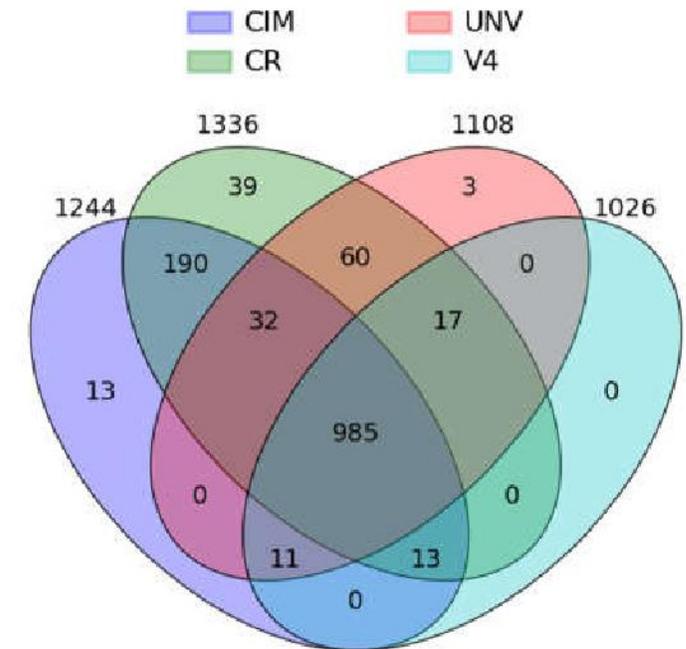


Figure 2: Répartition des séjours AVC au sein des différentes combinaisons des sources de repérage au CHU de Bordeaux.

# Grands enseignements

## Organisationnels

- Conventionnements complexes
- Budgets difficiles à prévoir  
100 000 euros par EDSH par indicateur/an/entrepôt
- Articulation entre acteurs :  
DSI / EDSH / équipes médicales

## Techniques

- Structuration et normalisation des données hétérogène
  - Dictionnaires et documentation des données indispensables mais rarement existants
- 
- Oriente le choix des indicateurs EDSH réalisables

## pour la mesure de la qualité

- Levier d'amélioration interne de la qualité
- Aide lors du développement de nouveaux indicateurs  
« dossier patient »
- Mesure ciblée sur les établissements avec un EDSH → nouveau type d'indicateur

# Détails sur l'étude AP-HP

Source : entrepôt de données de l'AP-HP

- Données du système d'information hospitalier
- Centralisées et standardisées aux formats OMOP et FHIR

## Tables utilisées

- Métadonnées d'imagerie (PACS) : structurées
- Diagnostics : structurées
- Données administratives des visites : structurées
- Données administratives des patients : structurées
- Comptes-rendus d'imagerie : non structurés

### Tables

FHIR	▼
IMAGING	>
IRIS	>
LIST / LIST_ENTRY	>
QUESTIONNAIRE	>
OMOP	▼
CARE_SITE	>
CONDITION_OCCURRENCE	
COST	
DRUG_EXPOSURE	>
LOCATION	>
MEASUREMENT	
NOTE	>
OBSERVATION	
PERSON	
PROCEDURE_OCCURRENCE	
PROVIDER	
VISIT_DETAIL	>
VISIT_OCCURRENCE	

# Cohorte initiale : outil cohort360

## Extraction depuis la base centrale

Un séjour en hospitalisation complète entre 2019 et 2023  
ET patient âgé 18 ans et plus lors d'au moins un séjour  
ET Un code diagnostic principal d'unité médical spécifique à l'AVC : I60, I61, I62, I63, I64

The screenshot displays the 'cohort360' tool interface for configuring a cohort. The breadcrumb path is 'IQSS AVC > première cohorte CIM10 > Version 34'. The 'POPULATION SOURCE' is set to 'APHP - ASSISTANCE PUBLIQUE AP-HP'. The query is structured as follows:

- ET 1**: Critère de prise en charge:
  - Âge : entre 18 an(s) et 130 an(s)
  - Prise en charge entre le 01/01/2019 et le 31/12/2023
  - Nombre d'occurrences >= 1
- OU 2**: Critère de diagnostic:
  - I60 - Hémorragie Sousarachnoïdienne - I61 - Hémorragie Intracérébrale - I62 - ...
  - Source: ORBIS
  - Prise en charge entre le 01/01/2019 et le 31/12/2023
  - Nombre d'occurrences >= 1
- OU 3**: Critère de diagnostic:
  - I60 - Hémorragie Sousarachnoïdienne - I61 - Hémorragie Intracérébrale - I62 - ...
  - Source: AREM
  - Prise en charge entre le 01/01/2019 et le 31/12/2023
  - Nombre d'occurrences >= 1

On the right side, there are control buttons: 'Créer la cohorte', 'Annuler', 'Rétablir', 'Réinitialiser', and 'Partager ma requête'. Below these, the 'ACCÈS' is 'Pseudonymisé' and 'PATIENTS INCLUS' is '49 428'. A 'VERSIONS DE LA REQUÊTE' section lists 34 versions, with the most recent being 'Version 34 - 18/06/2024 - 12:33:06'.

# Identification de la population cible

## Initial population

Critères cim-10 *grossiers* avec cohort-360

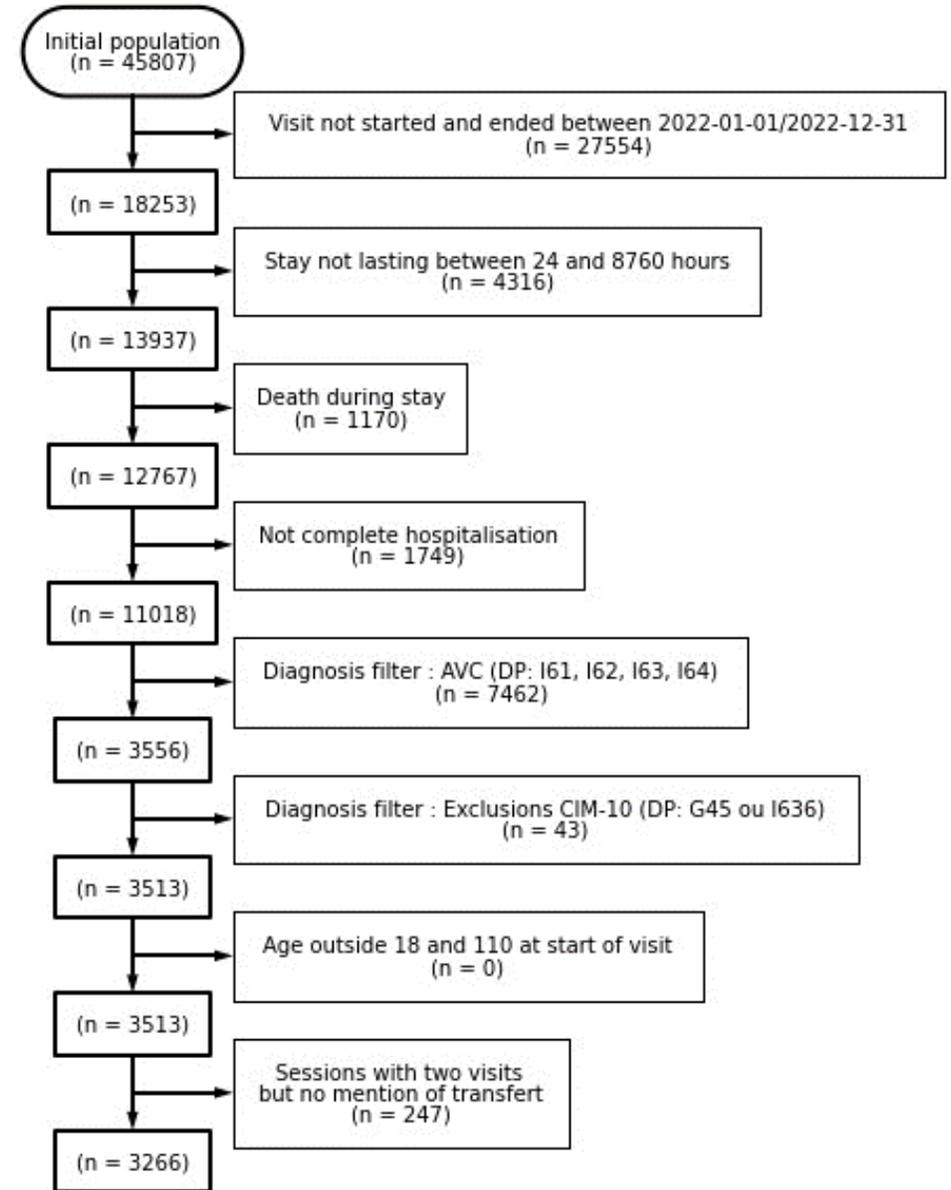
## Source CIM-10, 2022

Patients **âgés de 18 ans** ou plus lors du diagnostic,  
ET avec un **séjour** ayant commencé et pris fin en **2022**,  
ET une durée d'**hospitalisation >= 24h**,  
ET hospitalisation **complète**  
ET un **code CIM-10 AVC** (I61, I62, I63, I64) en DP de la  
première unité médicale visitée par le patient  
ET l'exception de I63.6 (Thrombose veineuse non  
pyrogène).

## Enfin,

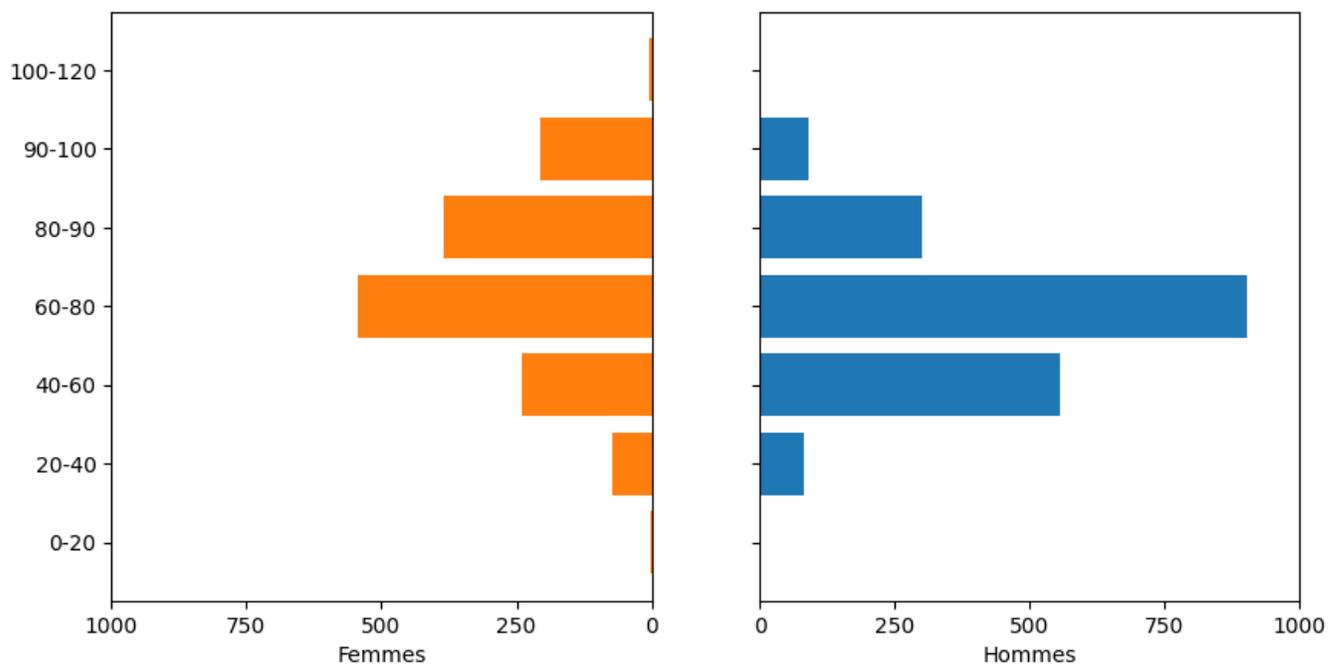
**Patients** inclus : 3266

**Séjours** inclus : 3390



Nombre de **patients** inclus

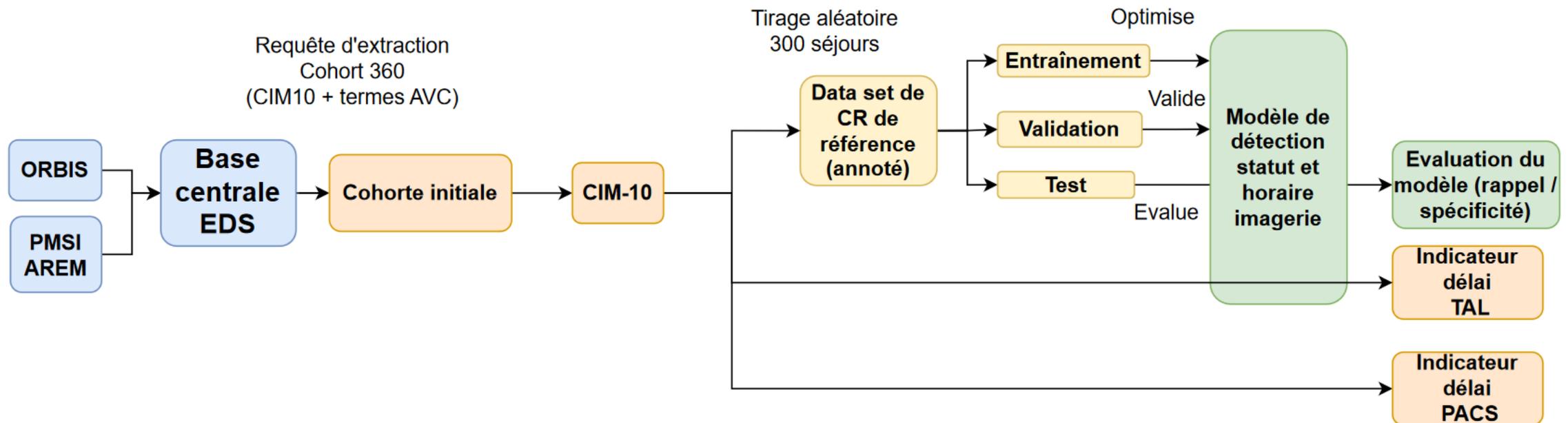
# Description de la population cible



Category	Variable	AVC_CIM-10
Demographics	Number of patients	3266.0
Demographics	Number of stays	3390.0
Demographics	Median age	71.61
Demographics	SD age	16.28
Demographics	Female ratio	0.43
Demographics	UNV/USINV ratio	0.48
Top care sites	Non renseigné	3257.0
Top care sites	HMN NEUROLOGIE	458.0
Top care sites	BCT UNV URG NEURO NPRGME (UF)	454.0
Top care sites	PSL URG CEREBRO VAS	427.0
Top care sites	BCT HC USINV	426.0
Top comorbidities	I10: Hypertension essentielle (primitive)	4570.0
Top comorbidities	G8100: Hémiplégie flasque récente, persistant au-delà de 24 heures	1064.0
Top comorbidities	R471: Dysarthrie et anarthrie	1020.0
Top comorbidities	G819: Hémiplégie, sans précision	987.0
Top comorbidities	R4700: Aphasie récente, persistant au-delà de 24 heures	966.0

# Deux approches différentes

- Méthode 1 : **Méta-données d'imagerie** : Les méta-données du PACS sont disponibles
- Méthode 2 : **Comptes-rendus** : Les délais sont renseignés dans les comptes-rendus d'imagerie (CRI)



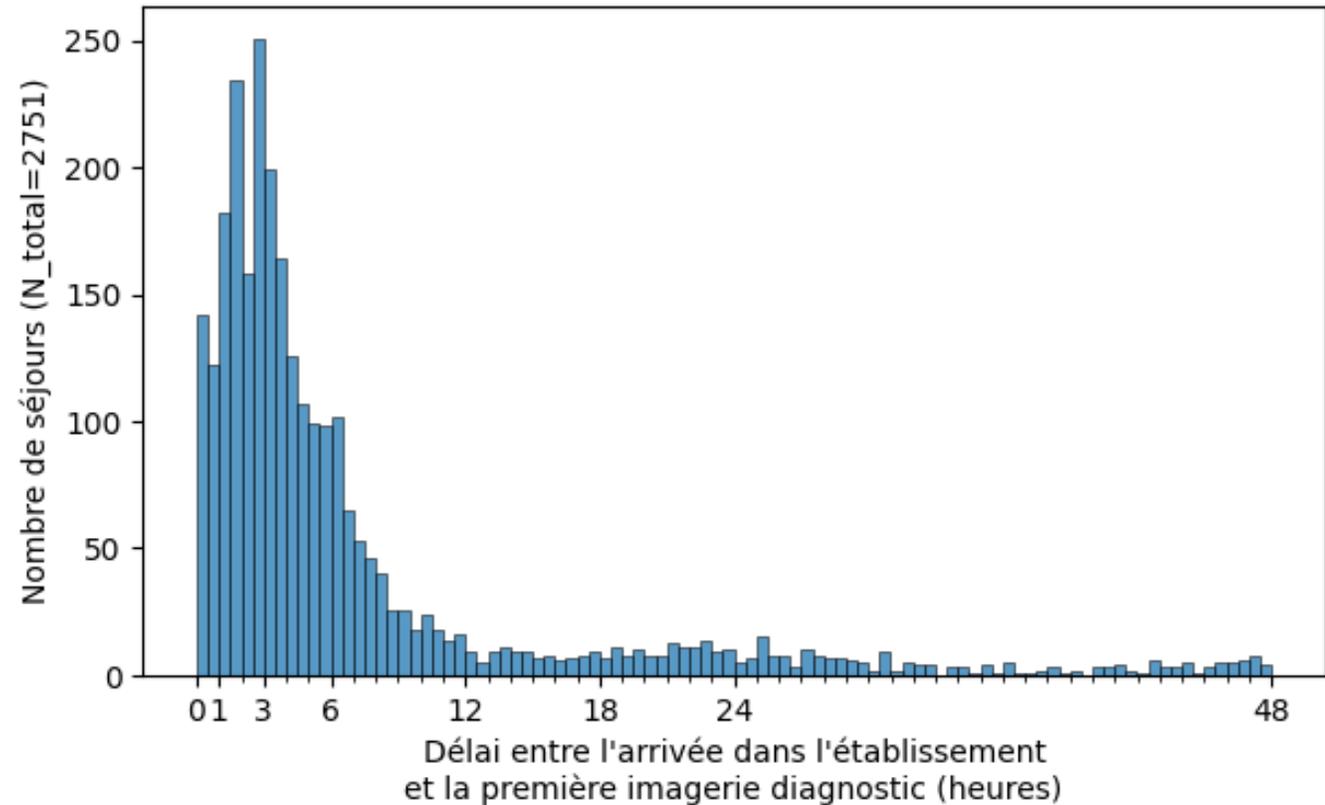
# Méthode 1 : Métadonnées du PACS

## Identification des imageries diagnostiques

- *Modalité normalisée* :  
[IRM, Tomodensimétrie]
- *ET Partie du corps normalisée* :  
[crâne, cerveau]
  - OU champ libre *Description*  
mentionnant le crâne ou le cerveau

### ⚠ Limites

Des métadonnées du PACS manquent dans l'EDS  
: 81% retrouvés (2751/3390 séjours)



# Méthode 2 : extraction depuis les comptes-rendus d'imagerie

Constat : Métadonnées des textes de mauvaise qualité

Objectif : Extraire pour chaque séjour,

- le statut de la 1<sup>ère</sup> imagerie diagnostique (IRM ou scanner)
- Les horodatages

## Méthode

- Annotation manuelle d'un sous-ensemble de 300 séjours
- Prompt un grand modèle de langage (LLM) validé sur 60 séjours
- Comparaison entre les deux approches sur 150 séjours

## Limites

- Parfois, seuls les horodatages de diction voire de validation du CR mentionnés
- Des CRI manquent : 74% retrouvés (2523/3390 séjours)

## Exemple (synthétique) de CRI

```
Otologie et Neuropsychologie
[Rosette](PRENOM) [Moreau](NOM)
Née le [4 septembre 1952](DATE_NAISSANCE), [1 60 04 25 311
-La Barre](VILLE).
0,5 1 1,5 2 2,5 3 3,5 4 4,5 5 5,5 6 6,5 7 7,5 8 8,5 9 9,5
Patient [105257992](IPP)
Compte-rendu hospitalisation du [01/07/2020](DATE) au [15/1
Mr [ROUSSEAU](PRENOM) [Pierre](NOM)
Né en [novembre 1975](DATE_NAISSANCE) [1 60 04 25 311 114 :
Hospitalisation courte dans le service d'ORL, CH [Henri Mo
[105-84 avenue du General de Gaulle](ADRESSE), [94010](ZIP
Résumé entrée :
Dès l'admission, une IRM cérébrale fut prescrite et réalis
Conclusion :
Présence d'un volumineux méningiome de l'angle ponto-céréb
Rendez-vous dans un mois avec le Dr [Chiraz](NOM) [Mohamad:
Imprimé en France par Ateliers d'Imprimerie, [13 rue Jeann
Date d'impression : [10 mars 2021](DATE) 10:30
Responsable de l'hôpital d'extraction : Dr. [Mohamed](PREN
Hôpital paris croix saint-simon [B. Duhamel](NOM) s/ servi
REDIGE PAR : Interne [Pierre](PRENOM) [ROUSSEAU](NOM) [rou:
```

# Annotation manuelle de 300 séjours parmi la population cible

## Annotateurs

- **3 annotateurs HAS** : 140 séjours par personne (2 à 3h / personne)
- Très bon accord inter-annotateur : Cohen-Kappa=0,87

## Annotations

- **Statut** de la première **imagerie** diagnostique (IRM, scanner...)
- **Date et l'horaire** de l'imagerie diagnostique
- Notion d'**imagerie antérieure**

## Logiciel d'annotation

Metanno : développé à l'AP-HP, open source, complet et simple

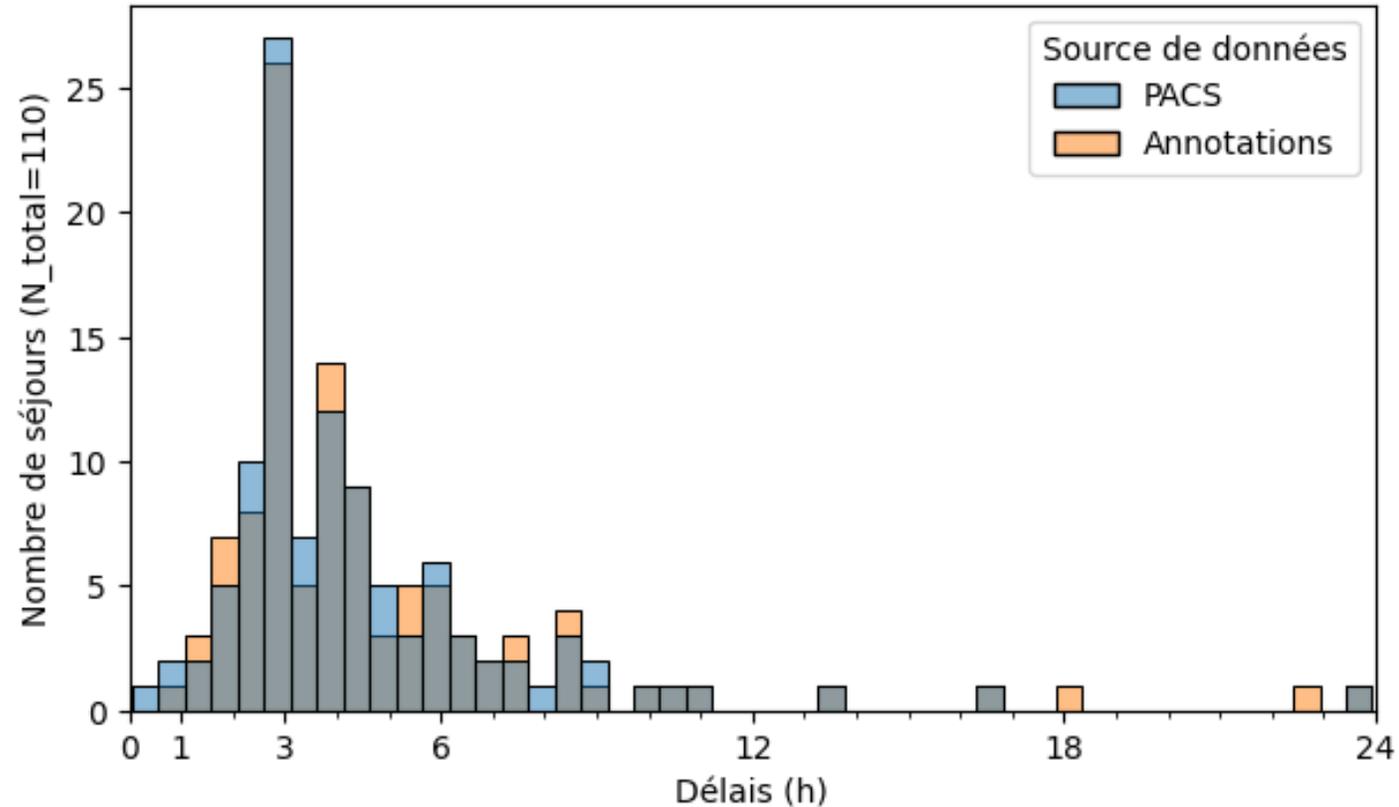
# Comparaison des délais entre le PACS et l'annotation

## Restrictions aux séjours :

- Sans notion d'imagerie antérieure (230/300)
- Pour les séjours avec un horaire dans le PACS ET dans les annotations (183/230)
- Uniquement avec un horaire d'examen annoté (110/183)

## Résultats

Pour 90% (99/110) de ces séjours, **différence de moins de 8 min** entre les deux horodatages.



# Comparaison entre l'annotation et le LLM

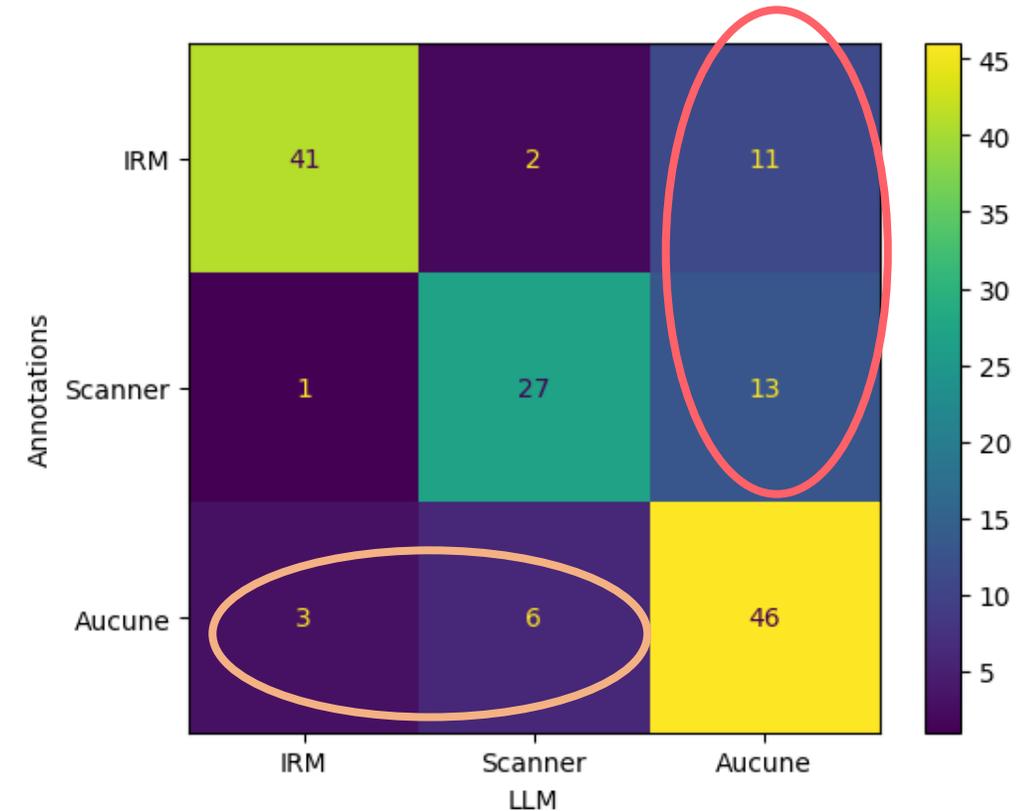
Quel est le type d'imagerie ?

Correspondance exacte = 76 % (118/150)

Principales erreurs

○ Le LLM prédit **trop souvent une imagerie antérieure ou un bilan**

○ Le LLM a du mal lorsqu'il y a un **angioscanner ET une imagerie à visée diagnostique (IRM ou scanner)**



Modèle : gemma 2 9B, 8-bit quantized

# Comparaison entre l'annotation et le LLM

Quels est l'horaire de l'examen ? (ou à défaut de diction ou de validation).

Restriction aux séjours valides

Séjours avec imagerie diagnostique et horodatage annoté lors de la campagne (95/150)

Correspondance exacte = 89 % (85/95)

count	mean	std	min	10%	25%	50%	75%	90%	max
95	-1274.08	12407.94	-120938.71	0	0	0	0	0	10.82

*Distribution des écarts en heure*

*(différences entre annotation et prédiction par gemma 2 9B, 8-bit quantized)*

## Principales erreurs

- Quelques hallucinations (horaire forcé à 00:00:00)
- Confusion avec l'horaire de début des symptômes



## IV. Conclusion et perspectives

# Enseignements pour les IQSS sur EDSH

Fort potentiel pour améliorer les pratiques

## Difficultés actuelles

- **Délais d'accès** aux données
  - **Manque de structuration** des données et de documentation
- La structuration et la normalisation actuelles des données des EDSH ne permet pas de calculer des indicateurs pour l'ensemble des thématiques de qualité des soins

# Absence du 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 pistes de solutions ?

 **Systematisation des appariements SNDS**

 **EDS de villes :** développement et articulation avec les EDSH

# Panorama des EDSH : bilan

## Investissements lourds

- Grands établissements uniquement ?
- Modèle économique à stabiliser

## Standards d'interopérabilité insuffisants

- **Mauvaise visibilité** de certaines informations pour les soignants
- **Pertes d'information** dans l'EDSH (limite les réutilisations secondaires)
- Difficultés de **standardisation** inter-établissements

## Gros efforts de standardisation a posteriori

- **Standardisation** dès la saisie dans le dossier
- Rôle de **l'ergonomie** des logiciels métiers, de l'IA pour l'aide à la saisie?

# Un modèle de financement en construction

## Vers un financement public pour l'investissement ?

- **Appel à projet HDH** (75M€ en 2023, moitié ONDAM, moitié France 2030) :  
16 lauréats couvrant l'ensemble des CHU
- **Recommandation d'une enveloppe annuelle** 60 à 90M€ (GT financement du HDH)
- Opportunité pour la **surveillance épidémiologique** : Projet Orchidée en cours par SPF

## Angle pertinence des soins

- Evaluation des **gains économiques** : ex. programme sSPREAD (redondance d'actes à Bordeaux)

## Inscription dans la politique nationale de qualité des soins

- **Rapport IGAS** sur la qualité des soins (Benoteau, Guillermo et Ricordeau, 2024)
- **Financement du temps médical** sous contrepartie de mesure locale de qualité

## Valorisation économique pour la mise à disposition

- Data Governance Act européen (2023)
- Cas d'usage identifié : pré-screening pour les essais cliniques

2023., Note à l'attention du Comité Stratégique du Système National des Données de Santé, financement des EDS au sein des CHU

# Perspectives

Au niveau national, diverses actions combinées en cours pour une exploitation plus aisée

- Contrat types pour accès aux données (DGE , ANS, HDH)
- Socle commun de données standardisées (HDH/EDS)
- Réseau d'entrepôts financé sur 4 ans pour des indicateurs épidémiologiques hospitaliers : Orchidée (SpF)

# Productions

## Rapport EDSH HAS 2022

[https://has-sante.fr/jcms/p\\_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france](https://has-sante.fr/jcms/p_3386123/fr/entrepots-de-donnees-de-sante-hospitaliers-en-france)

## Publication scientifique

Doutreligne, M., Degremont, A., Jachiet, P. A., Lamer, A., & Tannier, X. (2023). Good practices for clinical data warehouse implementation: A case study in France. PLOS Digital Health

<https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000298>

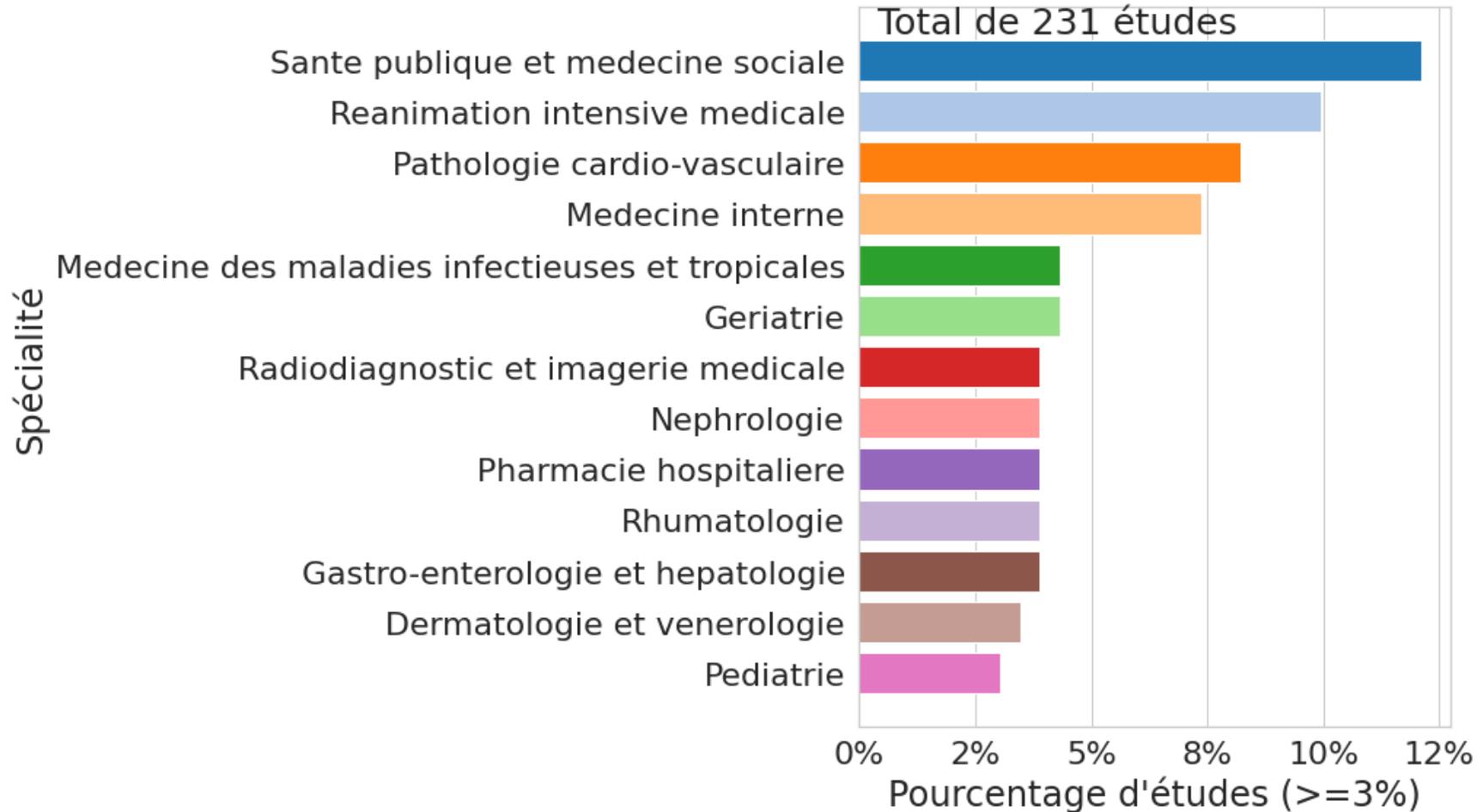
## Rapport de faisabilité IQSS sur EDSH (avec de nombreuses annexes techniques détaillant les études)

[https://www.has-sante.fr/jcms/p\\_3499689/fr/etude-de-la-faisabilite-et-de-l-interet-de-la-mesure-d-indicateurs-de-qualite-et-securite-des-soins-sur-les-entrepots-de-donnees-de-sante-hospitaliers](https://www.has-sante.fr/jcms/p_3499689/fr/etude-de-la-faisabilite-et-de-l-interet-de-la-mesure-d-indicateurs-de-qualite-et-securite-des-soins-sur-les-entrepots-de-donnees-de-sante-hospitaliers)



# Slides de supports

# Distribution des études par spécialité de l'investigateur principal



# Et ailleurs dans le monde ?

- German **Medical Informatics Initiative**, 2016
- English NHS funded **OpenSAFELY** platform, 2020
- US NIH funded **CHoRUS network**, 2022

# Choix des indicateurs

Listing de tous les IQSS mesurés à partir d'un audit de dossiers patients (74 MCO)

## Sélection d'IQSS :

- Sur la faisabilité : notion inspirée par *Fong et al., 2017 (AHRQ)* :
  - 1 concept, 1 critère
  - Temporalité, de 2 à 4 critères, plusieurs tables interrogées
  - Concepts polysémiques, réponses ouvertes, plus de 5 critères, plusieurs tables interrogées (complexe)
- Sur l'intérêt clinique : intérêt de disposer d'une mesure récente de l'indicateur
- Sur l'aspect stratégique

Proposition des IQSS pré-sélectionnés aux EDSH des CHU intéressés par la mesure d'indicateur

# Comptes-rendus de l'EDS AP-HP, Types de documents

Pour la population AVC CIM-10 en 2022, dans les 48 premières heures.

## Détails des titres :

- DOC-IMPOR : CR ECG (748), CR Samu/pompiers (225), Carte Groupe ABO (206), CR ECG, RPC NV (141)
- CR-IMAGE : Nonne (3756)
- INCONNU : Fiche Orthophonie (778), Fiche Kiné (627), None (481), fiche ergothérapie (239)
- CR-ACTE-DIAG-AUTRE : CR ECG (999), CR ECG;RPC NV (272), CR Acte diag. (214)
- CR-URGE : CR Urgences (259), puis CR Urgences pour chaque nom d'unité d'urgence de l'AP.

note_class_source_value	count
DOC-IMPOR	4725
CR-IMAGE	4251
INCONNU	3219
CR-ACTE-DIAG-AUTRE	2567
CR-URGE	2195
CRH-HOSPI	2067
CR-CONS	1619
LT-ORDO	1067
CR-SAMU	980
LT-AUTR	897
DOC-PAT	656

# Logiciel d'annotation :

Table des sessions

The screenshot displays the annotation software interface with the following components:

- Terminal 2:** Shows the file 'annotation\_matthieu.py'.
- Table des sessions:** A table with columns: person\_id, visit\_occurrence\_id, session\_id, and seen. The data is redacted with blue scribbles.
- Table des notes (filtrées dans la session sélectionnée):** A table with columns: note\_id, person\_id, visit\_occurrence, session\_id, note\_datetime, note\_class\_sourc, seen, and count. It shows two rows of data, with the first row highlighted.
- Table des annotations (filtrées dans la note sélectionnée):** A table with columns: offsets, delete, mention, and label. It shows one row of data.
- Output View:** Displays a list of terms: IRM, Scanner, date\_realisation\_imagerie, horaire\_realisation\_imagerie, date\_diction\_cr, horaire\_diction\_cr, date\_validation\_cr, horaire\_validation\_cr, notion\_imagerie\_anterieur, aucune\_notion\_AVG, and besoin\_aide.
- Texte de la note:** A text area containing a medical note starting with 'COMPTE RENDU D'EXAMEN' and discussing administrative, social, and medical data collection for patient care and management.

# Les thématiques d'intérêt

