HEALTH DATA HUB

Health Data Hub

La SNOMED-CT et usage secondaire des données de santé

Webinar ANS - 12/06/2025

Ordre du jour

A. **Le Health Data Hub**
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
    a. *CIM-10*
    b. *CCAM*
    c. *LPP*
E. Défis rencontrés

Le Health Data Hub a été impulsé et soutenu dans le cadre de politiques publiques d'ampleur

Les usages des données de santé se multiplient et accéder aux sources de données dans des délais les plus courts possibles est essentiel. Le Health Data Hub est une structure publique créée fin 2019 visant à faciliter l'accès aux données de santé aux projets d'intérêt public, dans la continuité de l'ouverture du SNDS en 2016.

Un **guichet unique** sur les données de Santé en France

Un **catalogue de données**, incluant une des plus large base des donnée médico administrative au monde

Une **plateforme sécurisée** et **à l'état de l'art**

En 5 ans, d'importants jalons ont été franchis

Mise en production d'une plateforme technologique en seulement un an

Recrutement de collaborateurs pour atteindre 116 personnes (janvier 2025)

Accompagnement de 176 projets dont un tiers portés par des acteurs industriels

Une implication européenne majeure, notamment leader du projet HealthData@EU chargé par la Commission européenne de **construire la première version d'EHDS**

Plus de 100 partenariats actifs concrétisant le développement de liens avec tout l'écosystème, y compris la société civile

Une impulsion majeure pour la constitution de nouvelles bases de données : AAP sur les entrepôts de données de santé (75M€), projet P4DP (10M€)

Plus de 650 interventions publiques depuis 2021 et **des participations dans de grands projets comme Parisanté Campus, FIAC, DARWIN**

176 projets accompagnés dont 131 utilisent la plateforme technologique

40 projets en phase réglementaire d'accès aux données de santé (125 nécessitent une aut. CNIL)

34 projets en phase de préparation des données

42 projets en fonctionnement sur la plateforme début 2025

23 projets terminés ou ayant produit des résultats intermédiaires

24 appels à projets à destination de l'écosystème

Projets du HDH avec OMOP terminés ou en cours

	Appels à projets EHDEN		Pilote EHDS - EMA Cas d'usage "Covid-19 et coagulopathies"	EMA - EMC2	EMA - DARWIN EU®
	EHDEN	**PERSEPHONE**			
Dates	• Début : 06-2020 • Fin des travaux : 12-2022	• Début : 12-2021 • Fin des travaux : 11-2023	• Début : 04-2024 • Livraison des données : 12-2024	• Début : 12-2021 • 1ère livraison des données : T3-2025	• Début : 01-2024 • 1ère livraison des données : T2-2025
Périmètre	• 100k patients avec diagnostic hospitalier de Covid-19 • **SNDS FT** 2019-2020 • **OMOP v5.3**	• **3M** patients • **Base principale du SNDS** 2015-2021 • **OMOP v5.3**	• **10M** patients • **Base principale du SNDS** 2017-2023 • **OMOP v5.4**	• **> 200k** patients / an • **BP du SNDS** 2015-2023 appariée aux données des patients de 4 hôpitaux puis actualisation et enrichissement annuel des données pdt 5 ans • **OMOP v5.4**	• **10M** patients • **BP du SNDS** 2016-2028 par périodes de 10 ans glissantes (9 ans + l'année en cours) avec actualisation annuelle • **OMOP v5.4** • **>15 projets / an attendus**

Ordre du jour

A. Le Health Data Hub

B. La standardisation des données : OMOP-CDM

C. Organisation des alignements de terminologie au HDH

D. Alignements SNOMED-CT

   a. *CIM-10*

   b. *CCAM*

   c. *LPP*

E. Défis rencontrés

Qu’est-ce que le modèle OMOP ?

Que veut dire OMOP-CDM ?

Observational **m**edical **o**utcomes **p**artnership - **C**ommon **D**ata Model.

C’est donc un **modèle de données standard.**

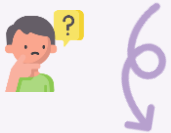
Standardisation syntaxique

Tables et variables standards

Standardisation sémantique

Vocabulaires standards (**SNOMED-CT**, Loinc, RxNorm, etc)

Pourquoi utiliser OMOP-CDM ?



Mener des études :

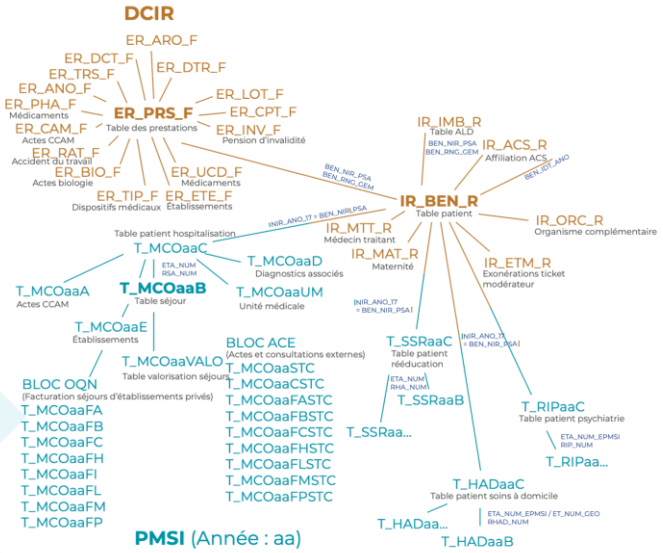
- **Observationnelles**
- **Fédérées**
- **A grande échelle**

⇒ scripts d’analyses standardisés

Utiliser OMOP-CDM pour la gestion des données interne d’un établissement n’est **PAS** vraiment adapté !
→ Risque de mettre des détails importants de côté

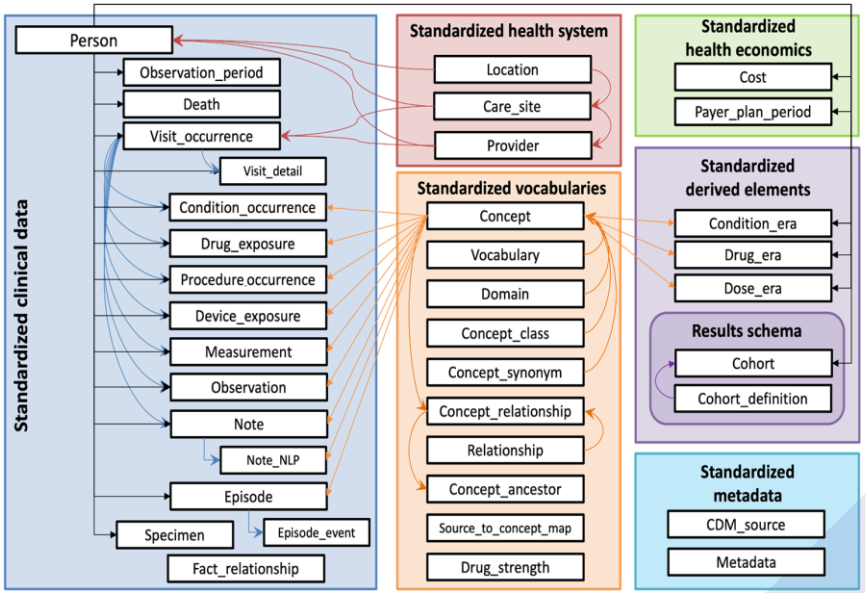
Transformation de la BP du SNDS au format OMOP v5.4

Base principale du SNDS



- **Schéma complexe** : **180 tables** - **4 500 variables**
- **Nombreuses terminologies** (CIM-10, CIP, CCAM, LPP, UCD, NABM, CSARR, GHM…)
- Centré sur **les soins**

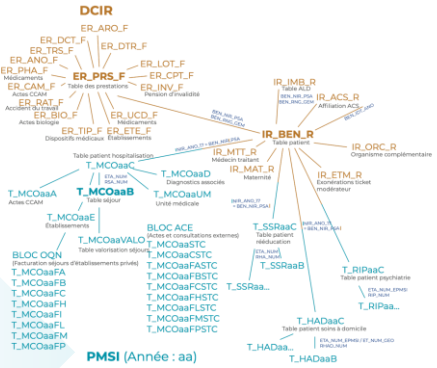
Données au format OMOP



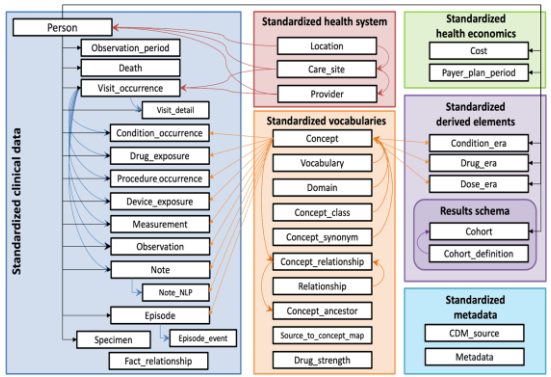
- Schéma centré autour de la **table PERSON**
- **29 tables** cible, vocabulaires standards
- **30 pays engagés** dans la standardisation vers OMOP-CDM en Europe (+130 sources de données)

Transformation de la base principale du SNDS au format OMOP-CDM

BP du SNDS historique



BP du SNDS OMOP-CDM



Collaborateurs du projet

Data engineer

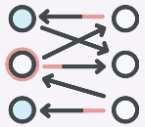
Internes en médecine / pharmacie

Data scientist

Experts SNDS

1. Établir une **correspondance** entre les champs (tables, variables) des deux formats
2. Écrire les **scripts** permettant de passer d'un format à l'autre
3. Vérifier la **qualité** de la base de données

OMOP-CDM utilise des terminologies standards



Toutes les **variables XX_concept_id** du modèle OMOP sont à **renseigner** selon des **concepts standards** issus de vocabulaires standards.

Tous les **concepts standards** sont disponibles sur le **portail Athena**. Une terminologie est définie comme standard par le groupe de travail sur les terminologies d'OHDSI.

Exemple de terminologies standards dans OMOP-CDM

Terminologies standards	Domaine(s) d'utilisation
SNOMED-CT	Procédure, pathologie, biologie, matériel médical, etc.
LOINC	Biologie
Gender	Genre du patient
Visit	Types de visites (urgence, visite à domicile, etc.)

Pourquoi parle-t-on d'alignements entre terminologies quand on fait de l'OMOPisation ?

Dans l'exemple du genre du patient:

- **Dans la base principale du SNDS** : 2 pour les femmes / 1 pour les hommes
- **Dans OMOP-CDM** : 8532 pour les femmes / 8507 pour les hommes

Aligner des terminologies revient donc à faire les **correspondances** entre **concept standard** du modèle OMOP et les **concepts de la base de données sources**.

Ordre du jour

A. Le Health Data Hub
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
    a. *CIM-10*
    b. *CCAM*
    c. *LPP*
E. Défis rencontrés

Organisation

Travail coordonné entre :

Suivi du processus : briefs, réunions, contrôle qualité

Société prestataire de transcodage

1. **Vérification** des pré-alignements et **Alignement de novo** des codes non pré-alignés ou rejetés

Internes (médecine / pharmacie)

2. **Revue des alignements** réalisés par le prestataire

Expert médical

3. **Validation finale** des alignements et conciliation/arbitrage en cas de discordances

Organisation

- **Prestataire de transcodage** *(Use & Share) :* (revue du) pré-alignement des codes (CCAM, LPP…) en recherchant :
  - la correspondance exacte : “exact match”
  - sinon, une correspondance “best fit” avec une catégorisation en :
    - “wider” = plus générale (préférée car plus largement applicable)
      Exemple : "HLHB001 - Biopsie hépatique transcutanée" → "252752005 - Liver biopsy" ou "75582008 - Diagnostic procedure on liver".
    - “narrower” = plus spécifique → **interdit !**

- **Validation** par les internes en médecine :
  - Tous les alignements proposés par la société de prestation sont revus par un interne.
  - Peut valider ou proposer un code alternatif.

- Rôle de l’**expert médical** du HDH
  - Pas d’intervention de l’expert si l’alignement est validé par l’interne au niveau le plus fin.
  - Intervention de l’expert HDH si :
    - *Désaccord entre prestataire et interne.*
    - *Plusieurs codes cibles sont possibles (wider…).*
    - *Aucune correspondance trouvée. Dans ces cas, l’expert :*
      - *Arbitre et désigne un “best fit”.*
      - *Peut organiser une réunion de conciliation.*

Stratégie d'alignement vers SNOMED - Fichier global

Résultat attendu : fichier global unique d'alignements avec l'ensemble des propositions et décisions prises ⇒ intégration table ***source_to_concept_map*** *dans les scripts de l'ETL OMOP*

1

Référentiels

- **LPP :** liste de codes (CNAM) = version 769 publiée le 13 mars 2024.
- **CCAM** : liste de codes utilisée =
  - analytique descriptive à usage PMSI 2024 de l'ATIH et v77
  - tarifaire de la CNAM (v78.10)
- **Listes de codes SNOMED** disponibles en ligne **:**
  - ATHENA avec codes *concept_id* associés
  - Edition nationale française listing SNOMED-FR 2024 publiée sur le serveur multi terminologie (SMT) de l'ANS
  - Édition internationale en anglais :
    - *SNOMED-INT 2024 (SMT de l'ANS)*
    - *explorateur des codes SNOMED (site SNOMED).*

2

Données source

- **SNDS** :
  - Codes LPP et CCAM les plus fréquents (80%) des données DCIR et MCO du SNDS en 2023

Stratégie d'alignement vers SNOMED - Outils

Outils pour réaliser les alignements

- **ATHENA :** outil OHDSI en ligne permettant l'alignement CIM-10 ↔ SNOMED pour récupérer les codes concept
- **USAGI :** outil d'aide à l'alignement développé par OHDSI qui compare des chaînes de caractères ⇒ va comparer les codes français traduits en anglais avec les autres codes utilisés à l'international (SNOMED CT, LOINC, etc) qu'il connaît
- **Outil d'alignement des codes LPP vers SNOMED** développé par le HDH spécifiquement pour les LPP *(cf diapo suivante)*

Pré-alignement des codes LPP vers SNOMED-CT

- **Outil développé par le HDH** *(Fei Gao - Experte SNDS et Data Scientist)*
  - Fondé sur le NLP et l'usage de thésaurus médicaux pour détecter les correspondances sémantiques entre libellés LPP et concepts SNOMED CT.

- **Fonctionnalités principales** de l'outil :
  - Nettoyage des libellés LPP
  - Inventaire manuel d'abréviations (extraites des libellés LPP en français).
  - Transcription des abréviations en langage complet, validée par un expert en documentation médicale.

- **Traduction FR → EN** :
  - Réalisée manuellement via DeepL.
  - Relecture médicale systématique lors de l'étape d'alignement.

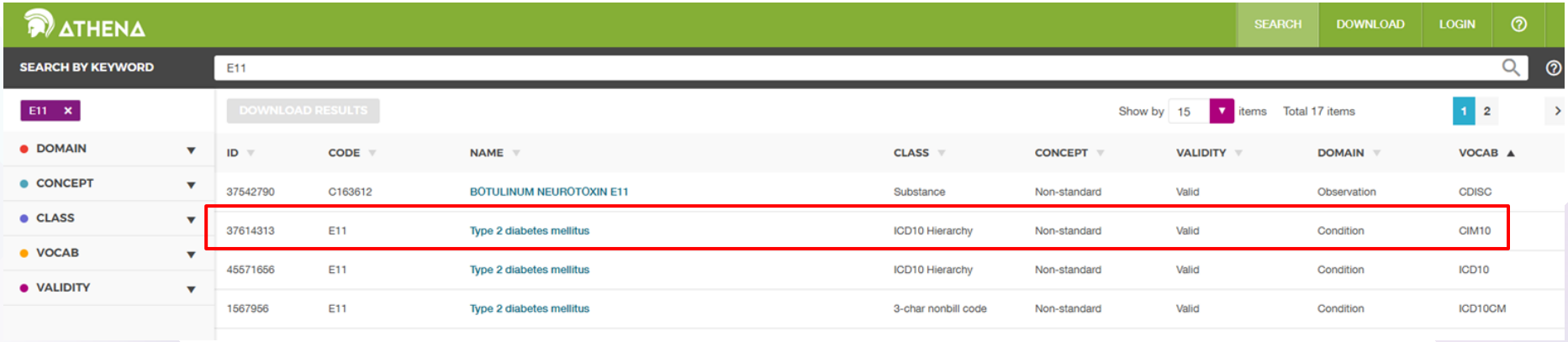
- **Pré-alignement** par similarité sémantique :
  - Calcul de scores de similarité entre le libellé LPP et les concepts SNOMED CT.
  - Proposition automatique des 10 concepts SNOMED CT les plus proches pour chaque code LPP.

Ordre du jour

A. Le Health Data Hub
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
   a. CIM-10
   b. CCAM
   c. LPP
E. Défis rencontrés

Outil d'alignement des codes CIM-10 vers SNOMED

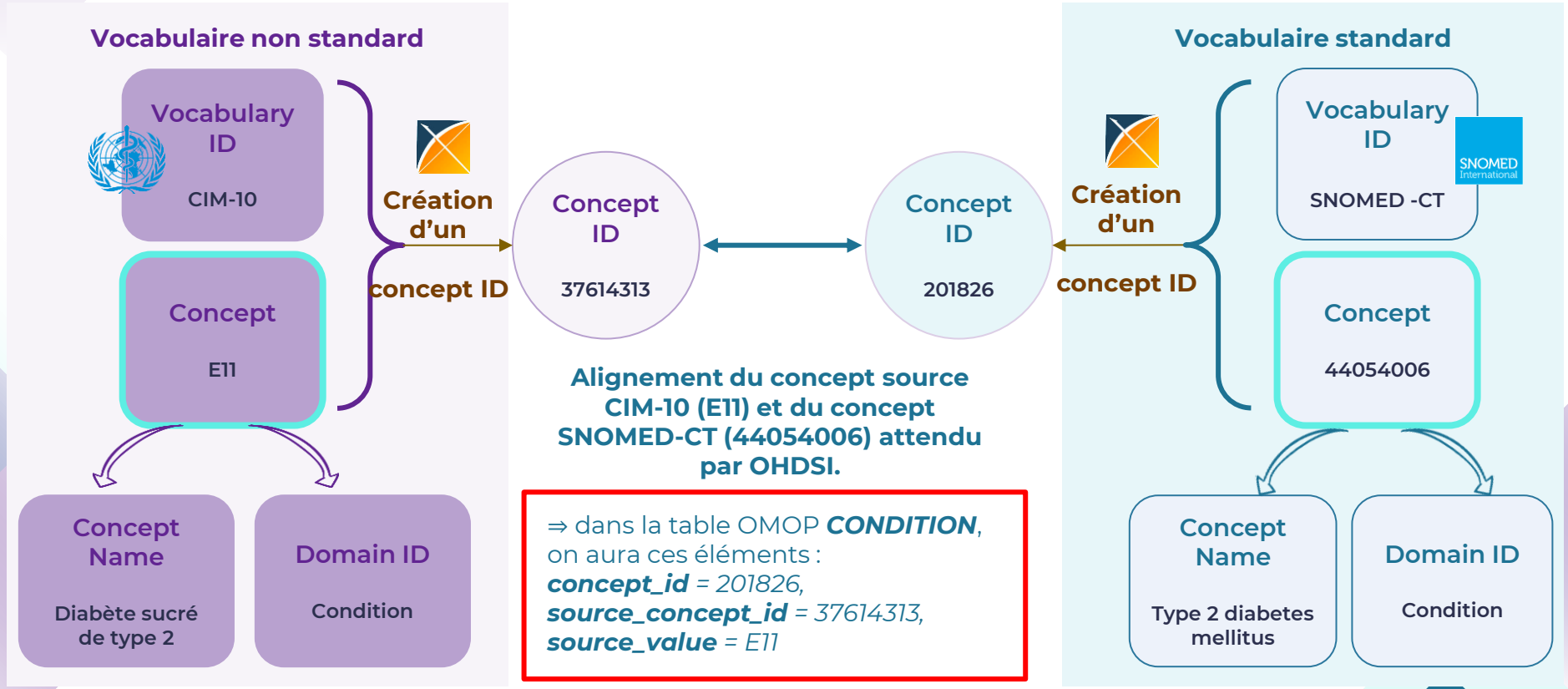
- **Outil ATHENA** (OHDSI) :
  - propose un **alignement direct** entre **codes CIM-10 / ICD-10** et **SNOMED** permettant de récupérer le *concept_id* associé qui figurera dans les tables OMOP
  - *Note : également alignement direct **CIP7 avec RxNorm** pour les médicaments mais pas pour CCAM, LPP, NABM ou CSARR*



Exemple :

- *le code CIM-10 E11 (Diabète de type 2) a pour concept_id "propre" 37614313 dans le standard OMOP*

Un exemple d'alignement pour un code CIM-10



Ordre du jour

A. Le Health Data Hub
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
   a. *CIM-10*
   b. *CCAM*
   c. *LPP*
E. Défis rencontrés

Constat sur les alignements CCAM à réaliser

- ~ **9 000** codes CCAM

- **Projets précédents** (EHDEN, PERSEPHONE) :
  - Alignements partiels ou trop généraux
  - Erreurs repérées (ex : concepts “narrower” non considérés comme tels)
  - Mise à jour SNOMED-CT et codes cibles “inactivés” (par catégorie, exemple : radiographie)

- Fréquence d’usage en 2023 (CNAM et ATIH)
  - **356 codes** couvrent 80 % des cas *(règle OMOP)*

- **Sur données du projet Darwin :**
  - **443 codes** couvrent 95% des occurrences
  - **139 validés,** 36 réalignés mais non arbitrés, 268 en cours d’alignement ⇒ **finalisé d’ici fin juillet 2025 (522 codes alignés en tout pour Darwin et EMC2)**

- **7 871 codes** moins fréquents : attribution du code SNOMED du niveau supérieur (sous-paragraphe, paragraphe, etc.) ⇒ **à revérifier mais non priorisé**.

Codes CCAM - Exemples d'alignements vers SNOMED

Exemple : pour le **sous-paragraphe CCAM** *01.01.01.01 "Électromyographie [EMG]"*:

- le sous-paragraphe lui-même est aligné avec la cible SNOMED *42803009 - "Electromyography" (concept_id = 4179815)*

- dans ce même sous-paragraphe :
  - le code principal *AHQB024* est fréquent et est aligné au niveau du code SNOMED cible *(252752005 "Concentric needle electromyography")*, l'alignement est intégré directement dans la table **source_to_concept_map** avec le concept_id associé (*4099522*)

  - le code principal *AHQB025* ne fait pas partie des codes fréquents et n'est pas aligné au niveau du code principal. Le code cible du sous-paragraphe *01.01.01.01.* lui sera attribué lors de l'intégration à la table **source_to_concept_map**

Ordre du jour

A. Le Health Data Hub
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
    a. *CIM-10*
    b. *CCAM*
    c. *LPP*
E. Défis rencontrés

Stratégie d'alignement des codes LPP vers SNOMED-CT

En entrée : le code source ainsi que son libellé en français

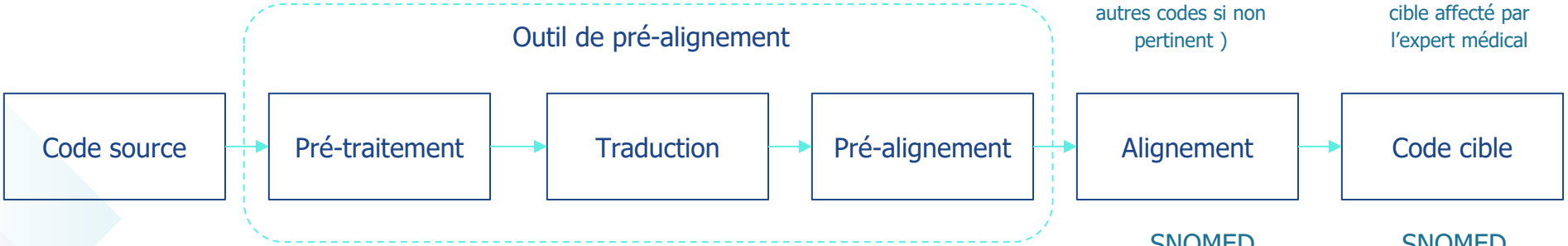
Pré-traitement du libellé (nettoyage, abréviations, règles métiers etc.)

Traduction du libellé en anglais via DeepL

N propositions de codes cibles sont fournies par l'outil avec les scores associés

Le premier expert médical choisit le code cible parmi les propositions fournies (ou recherche parmi les autres codes si non pertinent)

Le second expert médical valide le choix d'alignement final.
Le code source est aligné vers le code cible affecté par l'expert médical



LPP = 1333695
Pansement en fibres de cmc, >ou= 63 cm2 et < 100 cm2, boite de 10

1333695
Pansement en fibres de carboxyméthylcellulose.

1333695
carboxymethylcellulose fiber dressing.

1333695
carboxymethylcellulose fiber dressing.

SNOMED 467903007
concept = 45761673
Cotton burn dressing

SNOMED 467903007
concept = 45761673
Cotton burn dressing

4233400
Absorbent cellulose dressing with fluid repellent backing
P = 0.910496

4219809
Collagen dressing
P = 0.910602

45761673
Cotton burn dressing
P = 0.910865

Stratégie d'alignement des codes LPP vers SNOMED-CT

- **Liste LPP** : ~ 33 000 codes
  - **1 313 codes "génériques"** *(ex. "1396655 - Boîtes de 100 compresses non tissées non stériles de 100 cm²")*
  - **~ 31 700 codes "non génériques"** (avec marque, *ex. "6318832 - Boîtes de 100 compresses non tissées non stériles de 100 cm², Hartmann", "6318803 - Boîtes de 100 compresses non tissées non stériles de 100 cm², Raffin")*
    - ~ 23 100 codes (70%) sont rattachés aux 1 313 "génériques"
    - ~ 9 900 codes sont "non génériques" isolés (sans lien avec un "générique") dont 1 500 n'ont pas de libellé

- **Sur données du projet Darwin :**
  - **791 codes** couvrent 95% des occurrences (SNDS)
  - **533 codes** ont été pré-alignés (NLP) et revus par presta/internes, 258 restent à aligner ⇒ **finalisé d'ici fin juillet 2025**

- **Codes isolés** à traiter en phase 2 :
  - ~7 500 codes "non génériques" sans "générique" → alignement individuel **mais non priorisé**

Ordre du jour

A. Le Health Data Hub
B. La standardisation des données : OMOP-CDM
C. Organisation des alignements de terminologie au HDH
D. Alignements SNOMED-CT
    a. *CIM-10*
    b. *CCAM*
    c. *LPP*
E. Défis rencontrés

Défis rencontrés

- **Expertise médicale** nécessaire pour procéder à l'alignement car certains cas peuvent s'avérer complexes et il faut une connaissance des termes médicaux
- **Structure différente** : hiérarchie monoaxiale (CCAM, LPP) vs ontologie multiaxiale (SNOMED) → difficulté à trouver une correspondance directe (code "*wider*" privilégié)
- **CCAM/LPP :** un code unique peut couvrir plusieurs gestes médico-techniques ou dispositifs médicaux distincts, chacun ayant une cible SNOMED différente → arbitrages complexes
- **CCAM :** un même code principal est utilisé pour décrire une chirurgie et l'acte d'anesthésie associé ⇒ pas de prise en compte de l'activité associée dans l'alignement → empêche le repérage de l'activité d'anesthésie
- **LPP :** des informations présentes dans le libellé LPP sont inutiles pour l'alignement. A l'inverse, il est nécessaire d'aller rechercher la documentation de chaque code pour trouver certaines informations utiles (exemple : "stérile" pour un pansement)
- **SNOMED** : inactivation de certains codes lors des mises à jour annuelles
- **Dans tous les cas** : la traduction FR ⇒ EN est une étape critique.

Prochaines étapes

- **Finaliser les alignements EMC2 / Darwin :** d'ici fin juillet 2025
- **Darwin** : 1ere études attendues en septembre 2025
- **EMC2** : 1ères données EDS reçues en juillet 2025 (CLB)
- **Prise de contact avec OHDSI, ATIH, ANS :** discuter de la maintenance, mise à jour et ouverture des listes CCAM et LPP créées

Remerciements

- **Health Data Hub :** Claire Galland, Axelle Menu, Amelia Deguilhem, Fei Gao, Mathieu Goujeon, Constance Daucy, Victor Leandre-Chevalier
- **Use&Share et Earlytracks :** Anne Maheust, François Macary, Brice de Behaust, Charley Guillaume

Avez-vous des questions ?

HEALTH
DATA HUB

Suivez-nous sur les réseaux sociaux !