



HEALTH DATA HUB

Schémas d'appariement pour des projets réalisés au sein du Health Data Hub

04.2020

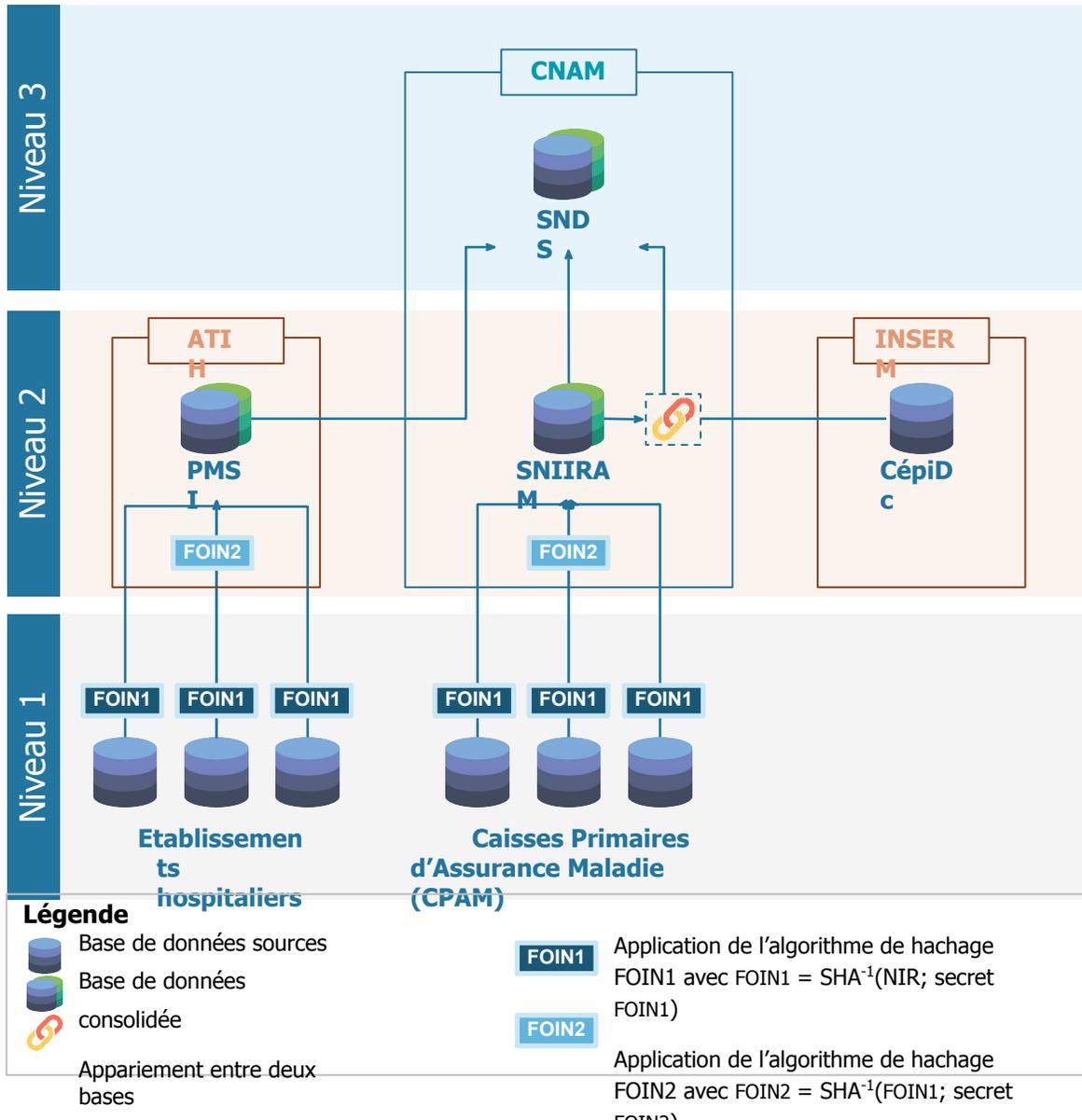
1

**Rappel sur le
SNDS**

2

**Circulation et appariement des
données**

Rappel sur le SNDS - périmètre historique (1/2)



Le SNDS est alimenté par trois bases de données :

- **SNIRAM**

Les Caisses Primaires d'Assurance Maladie remontent l'ensemble des informations issues des remboursements à la CNAM. Pour garantir la sécurité de ces données, la solution CAMEL est utilisée permettant le contrôle, le chiffrement, et le transfert sécurisé des flux vers le SNIIRAM. Le flux de données provenant des caisses passe par CAMEL qui applique la fonction de hachage FOIN1 sur le NIR.. Les données sont envoyées à la CNAM qui applique ensuite l'algorithme de hachage FOIN2.

- **PMSI**

Chaque établissement de santé enregistre chacun des séjours hospitaliers, appelé résumé de sortie standardisé (RSS) qui sont classés dans des Groupe Homogène de Malade (GHM) selon des critères médicaux et économiques. Les établissements utilisent alors le logiciel MAGIC qui applique la fonction de hachage FOIN1 sur le NIR et dépose les données pseudonymisées sur la plateforme e-PMSI de l'ATIH.

L'ATIH applique alors l'algorithme de hachage FOIN2 et envoie les données à la CNAM pour intégration dans le SNDS.

- **CépîDc**

Le CépîDc de l'Inserm gère la Base de Causes Médicales de Décès (BCMD). Il ne traite que la partie médicale du certificat de décès, par conséquent, la base ne contient aucun nom. Les données sont transmises à la CNAM pour intégration dans le SNDS et sont appariées de façon probabiliste au SNIIRAM pour récupérer les FOIN2 correspondants afin de lier les causes de décès aux autres données présentes dans le SNDS.

Rappel sur le SNDS - périmètre historique (2/2)

Il couvre l'ensemble des personnes ayant eu **recours au système de soins français ou étant décédées sur le territoire**. Il permet de disposer d'informations sur le recours aux soins, le montant des dépenses associées, les informations concernant les **séjours hospitaliers depuis 2006**, ainsi que les **causes de décès des individus sur la période 2013-2015**.

Les **principales informations du SNDS** sont :

- Informations sur le bénéficiaire (sexe, mois et année de naissance, rang de naissance, lieu de résidence, régime, couverture maladie universelle complémentaire, aide à la complémentaire santé) ;
- Informations sur les professionnels de santé (spécialité, mode d'exercice, sexe, âge, département d'implantation)
- Pathologies, notamment les affections de longue durée et les diagnostics des séjours hospitaliers ;
- Dépenses et remboursements (prestations en soins de ville, en établissements de santé, et montants associés)
- Consommations de soins de ville (consultations, actes techniques...)
- Prescriptions (médicaments)
- Dispositifs médicaux (aides techniques)
- Autres prestations (cures, transports...)
- Soins hospitaliers (hors séances)
- Séjours hospitaliers
- Indemnités journalières (maladie, ATMP, maternité) et invalidité
- Causes médicales de décès

Pour approfondir

- Dictionnaire des variables : [lien](#)
- Documentation du SNDS : [lien](#)
- Forum d'entraide : [lien](#)
- Contribuer (Gitlab) : [lien](#)
- Programme de formation – à venir

1

**Rappel sur le
SNDS**

2

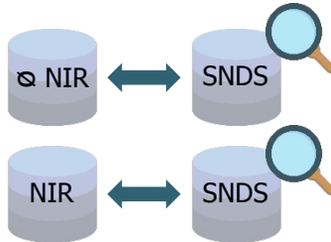
**Circulation et appariement des
données**

Présentation des différents types d'appariement

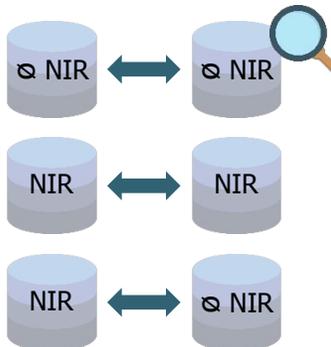
1. Appariements déterministes directs

Couplage de données réalisé à partir de champs strictement identifiants et communs aux deux sources de données (ex : NIR)

Appariement déterministe direct entre une base de données et le SNDS



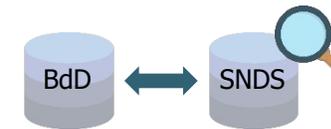
Appariement déterministe direct entre deux base de données (hors SNDS)



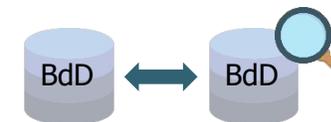
2. Appariements déterministes indirects

Couplage statistique de données réalisé à partir d'un ensemble de variables discriminantes et communes aux deux jeux de données

Appariement déterministe direct entre une base de données et le SNDS



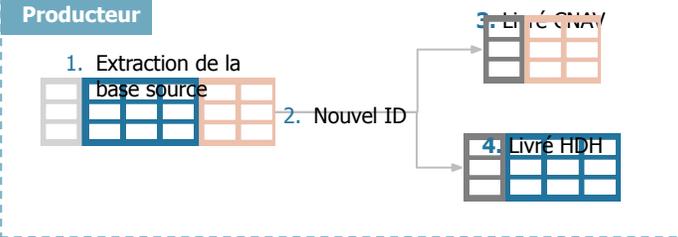
Appariement déterministe indirect entre deux base de données (hors SNDS)



Appariement déterministe direct entre une base de données sans NIR et le SNDS

#1 Extraction des données de la cohorte

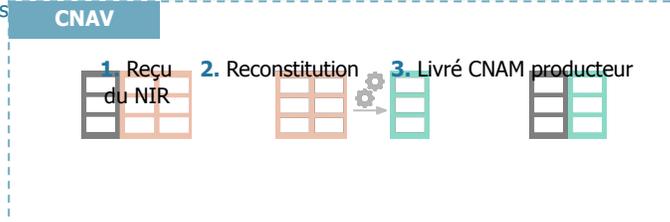
- Le producteur réalise une extraction de sa base source, avec :
 - L'ID interne : ID_source
 - Les traits identifiants : noms, sexe, date et lieu de naissance
 - Les données de recherche
- Un nouvel ID est généré aléatoirement : ID_temp
- Une table d'appariement est envoyée à la CNAV avec :
 - ID_temp
 - Les traits identifiants
- L'extraction de la base source est envoyée au HDH avec :
 - ID_temp
 - Les données de recherche pseudonymisées



Livraison des traits identifiants

#2 Recherche des Numéros d'Identification au Répertoire (NIR)

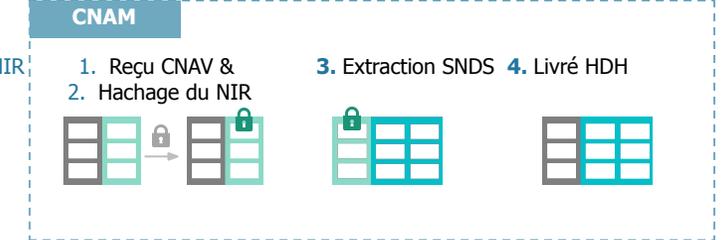
- La CNAV reçoit du producteur la table d'appariement, avec :
 - ID_temp
 - Les traits identifiants : noms, sexe, date et lieu de naissance
- La CNAV cible les NIR les plus approchants à partir des traits identifiants
- La table d'appariement mise à jour est envoyée à la CNAM, avec :
 - ID_temp
 - Le NIR (en clair)
 - Les traits identifiants : sexe et date de naissance uniquement
- Après envoi de la table d'appariement, les données relatives au nom, prénoms et lieux de naissance sont supprimées par la CNAV



Livraison des NIR

#3 Extraction des données SNDS

- La CNAM reçoit de la CNAV la table d'appariement, avec :
 - ID_temp
 - Le NIR (en clair)
 - Les traits identifiants : sexe et date de naissance uniquement
- La CNAM retrouve l'ID_SNDS grâce aux algorithmes de hachage du NIR (Foin 1 et 2)
- La CNAM cible les données du SNDS requises via l'ID_SNDS
- L'extraction du SNDS est envoyée au HDH, avec :
 - ID_temp
 - Données du SNDS requises (pseudonymisées)

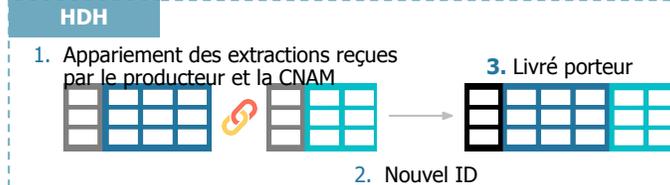


Livraison des données SNDS

Livraison des données de la cohorte

#4 Appariement des données

- Le HDH reçoit les extractions transmises par le producteur et la CNAM. Il les apparie à partir de l'identifiant ID_temp.
 - Un nouvel ID est généré aléatoirement : ID_projet
 - Un jeu de données final est livré à l'équipe projet, avec :
 - ID_projet
 - Les données de recherche pseudonymisées
 - Les données SNDS pseudonymisées
- Après validation de l'appariement et livraison du jeu de données, l'ID_temp et les tables d'appariement détenues par le producteur, la CNAV, la CNAM et le HDH sont supprimés



Légende

Identifiants

- ID_source (gris)
- ID_temp (gris foncé)
- ID_projet (noir)

Aléatoires

NIR

- NIR en clair (vert clair)
- ID_SNDS (NIR foïnisé) (vert foncé avec cadenas)

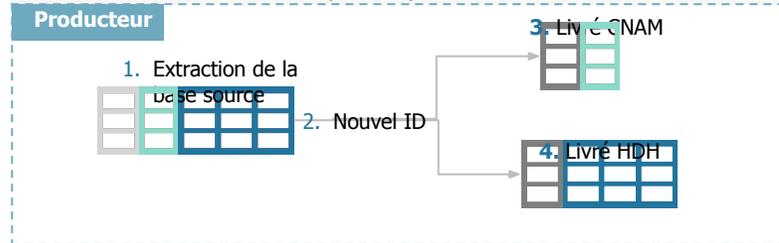
Données

- Données cohorte (bleu foncé)
- Traits identifiants (orange)
- Données SNDS (bleu clair)

Appariement déterministe direct entre une base de données avec NIR et le SNDS

#1 Extraction des données de la cohorte

- Le producteur réalise une extraction de sa base source, avec :
 - L'ID interne : ID_source
 - Le NIR (en clair), avec sexe et date de naissance
 - Les données de recherche
- Un nouvel ID est généré aléatoirement : ID_temp
- Une table d'appariement est envoyée à la CNAM avec :
 - ID_temp
 - Les traits identifiants
- L'extraction de la base source est envoyée au HDH avec :
 - ID_temp
 - Les données de recherche pseudonymisées



Livraison des NIR

#2 Extraction des données SNDS

- La CNAM reçoit du producteur la table d'appariement, avec :
 - ID_temp
 - Le NIR (en clair), avec sexe et date de naissance uniquement
- La CNAM retrouve l'ID_SNDS grâce aux algorithmes de hachage du NIR (Foin 1 et 2)
- La CNAM cible les données du SNDS requises via l'ID_SNDS
- L'extraction du SNDS est envoyée au HDH, avec :
 - ID_temp
 - Données du SNDS requises (pseudonymisées)

CNAM

- Reçu producteur & 3. Extraction SNDS
- Hachage du NIR
- Livré HDH



#3 Appariement des données

- Le HDH reçoit les extractions transmises par le producteur et la CNAM. Il les apparie à partir de l'identifiant ID_temp.
- Un nouvel ID est généré aléatoirement : ID_projet
- Un jeu de données final est livré à l'équipe projet, avec :
 - ID_projet
 - Les données de recherche pseudonymisées
 - Les données SNDS pseudonymisées

Après validation de l'appariement et livraison du jeu de données, l'ID_temp et les tables d'appariement détenues par le producteur, la CNAM et le HDH sont supprimés

HDH

- Appariement des extractions reçues par le producteur et la CNAM
- Nouvel ID
- Livré porteur



Livraison des données de la cohorte

Livraison des données SNDS

Légende

Identifiants

- ID_source
 - ID_temp
 - ID_projet
- Aléatoires

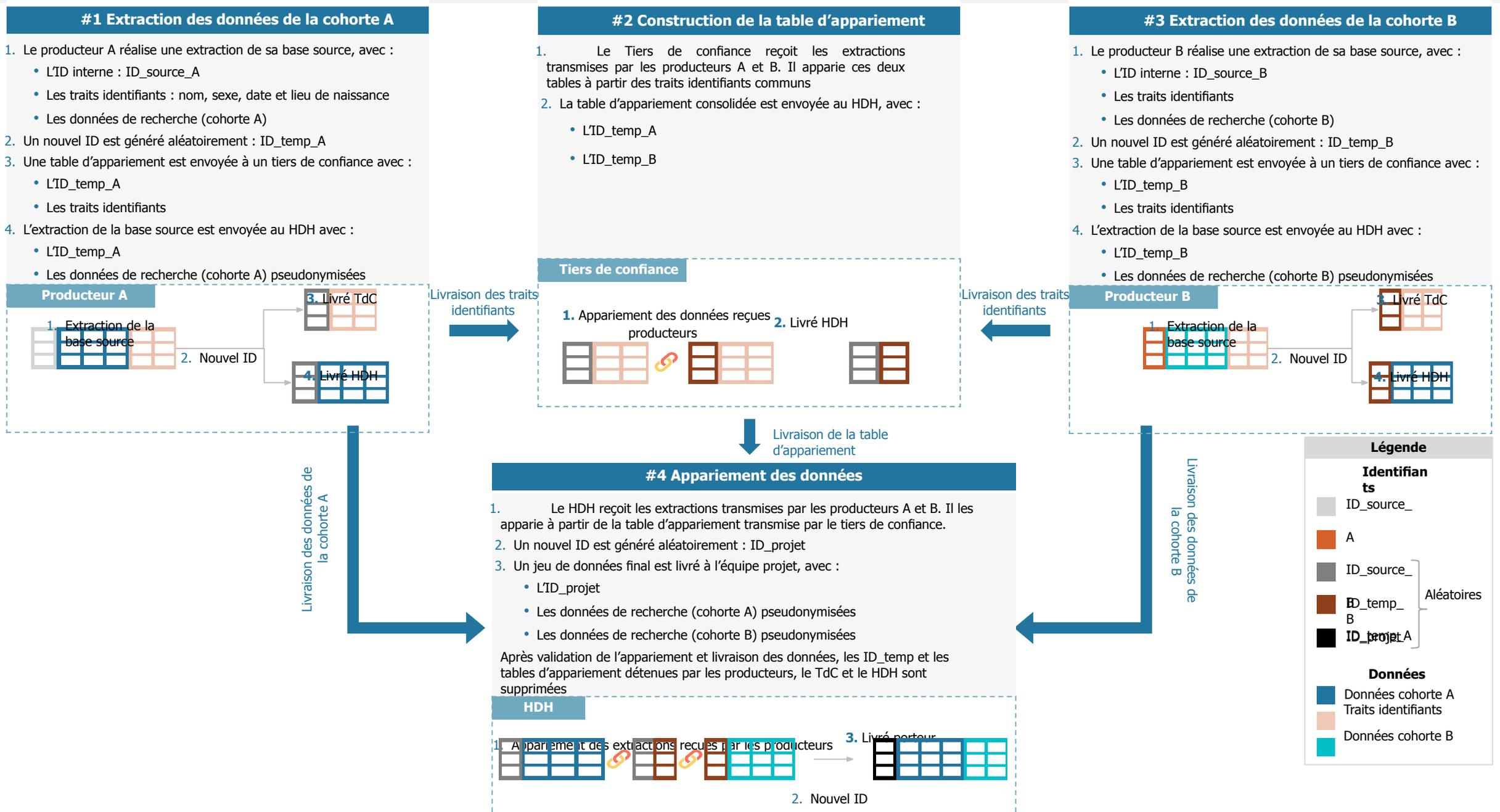
NIR

- NIR en clair, sexe, DDN
- ID_SNDS (NIR foinisé)

Données

- Données cohorte
- Données SNDS

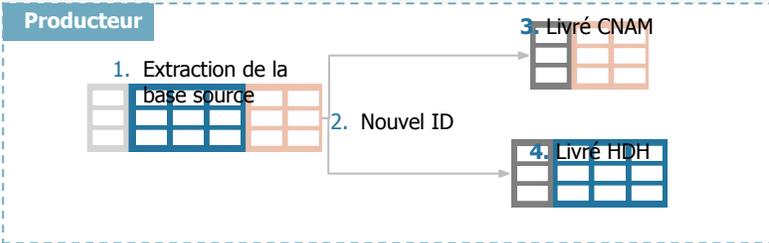
Appariement déterministe direct entre deux bases de données



Appariement déterministe indirect entre une base de données et le SNDS

#1 Extraction des données de la cohorte

- Le producteur réalise une extraction de sa base source, avec :
 - L'ID interne : ID_source
 - Les variables communes
 - Les données de recherche
- Un nouvel ID est généré aléatoirement : ID_temp
- Une table d'appariement est envoyée à la CNAM avec :
 - ID_temp
 - Les variables communes
- L'extraction de la base source est envoyée au HDH avec :
 - ID_temp
 - Les données de recherche pseudonymisées



Livraison de la table d'appariement

#2 Extraction des données SNDS

- La CNAM reçoit du producteur la table d'appariement, avec :
 - ID_temp
 - Les variables communes
- À partir des variables communes, la CNAM effectue une recherche probabiliste des patients dans le SNDS et cible les données pertinentes.
- L'extraction du SNDS est envoyée au HDH, avec :
 - ID_temp
 - Données du SNDS requises (pseudonymisées), pouvant comprendre une partie des variables communes

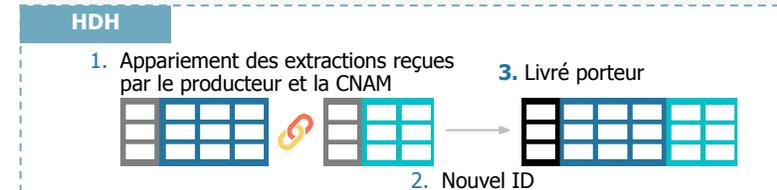


#3 Appariement des données

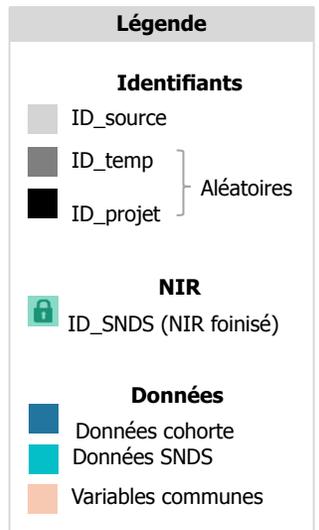
- Le HDH reçoit les extractions transmises par le producteur et la CNAM. Il les apparie à partir de l'identifiant ID_temp.
- Un nouvel ID est généré aléatoirement : ID_projet
- Un jeu de données final est livré à l'équipe projet, avec :
 - ID_projet
 - Les données de recherche pseudonymisées
 - Les données SNDS pseudonymisées

Après validation de l'appariement et livraison du jeu de données, l'ID_temp et les tables d'appariement détenues par le producteur, la CNAM et le HDH sont supprimés

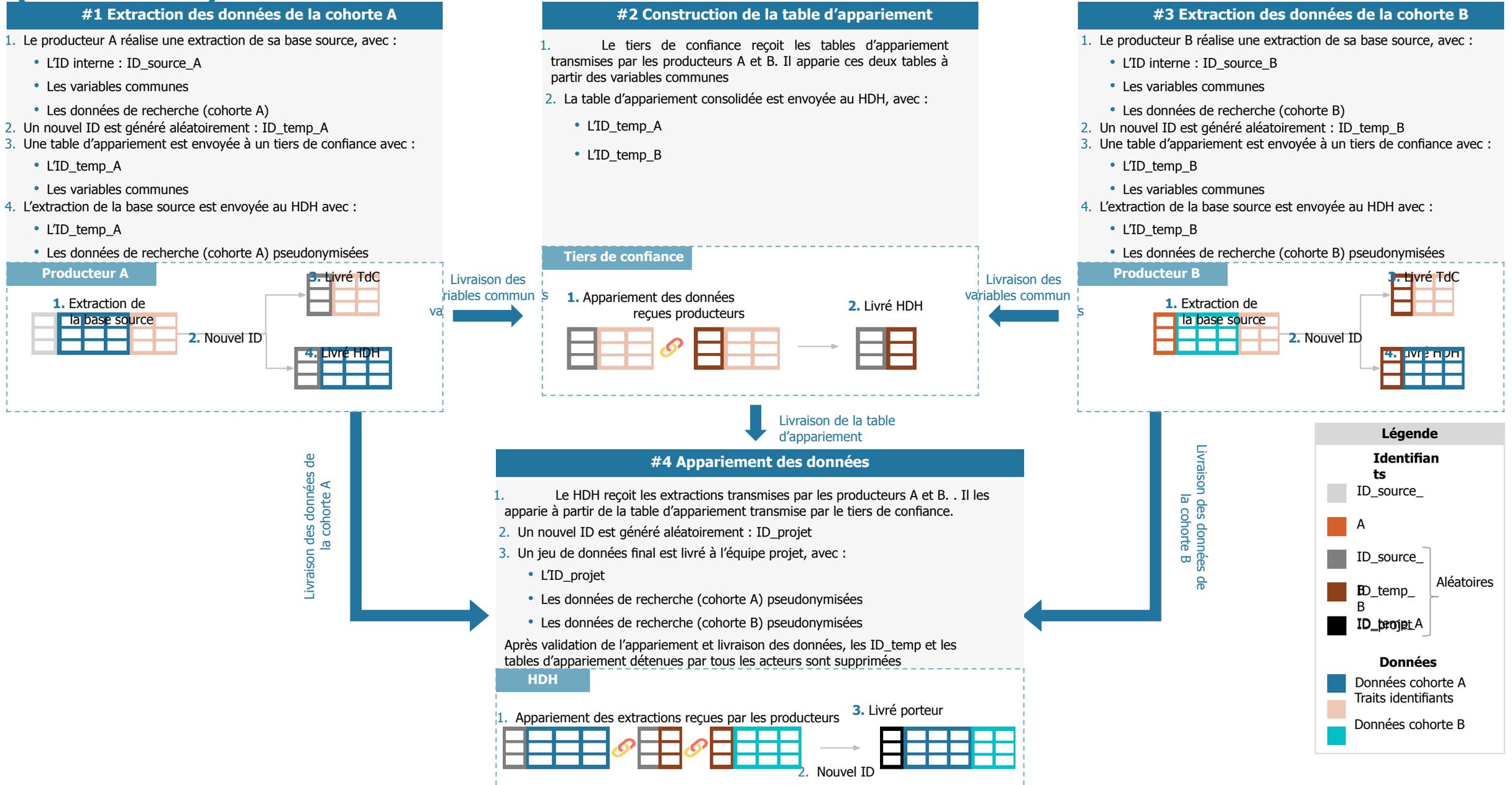
Livraison des données de la cohorte



Livraison des données SNDS



Appariement déterministe indirect entre deux bases de données (hors SNDS)





Health Data Hub