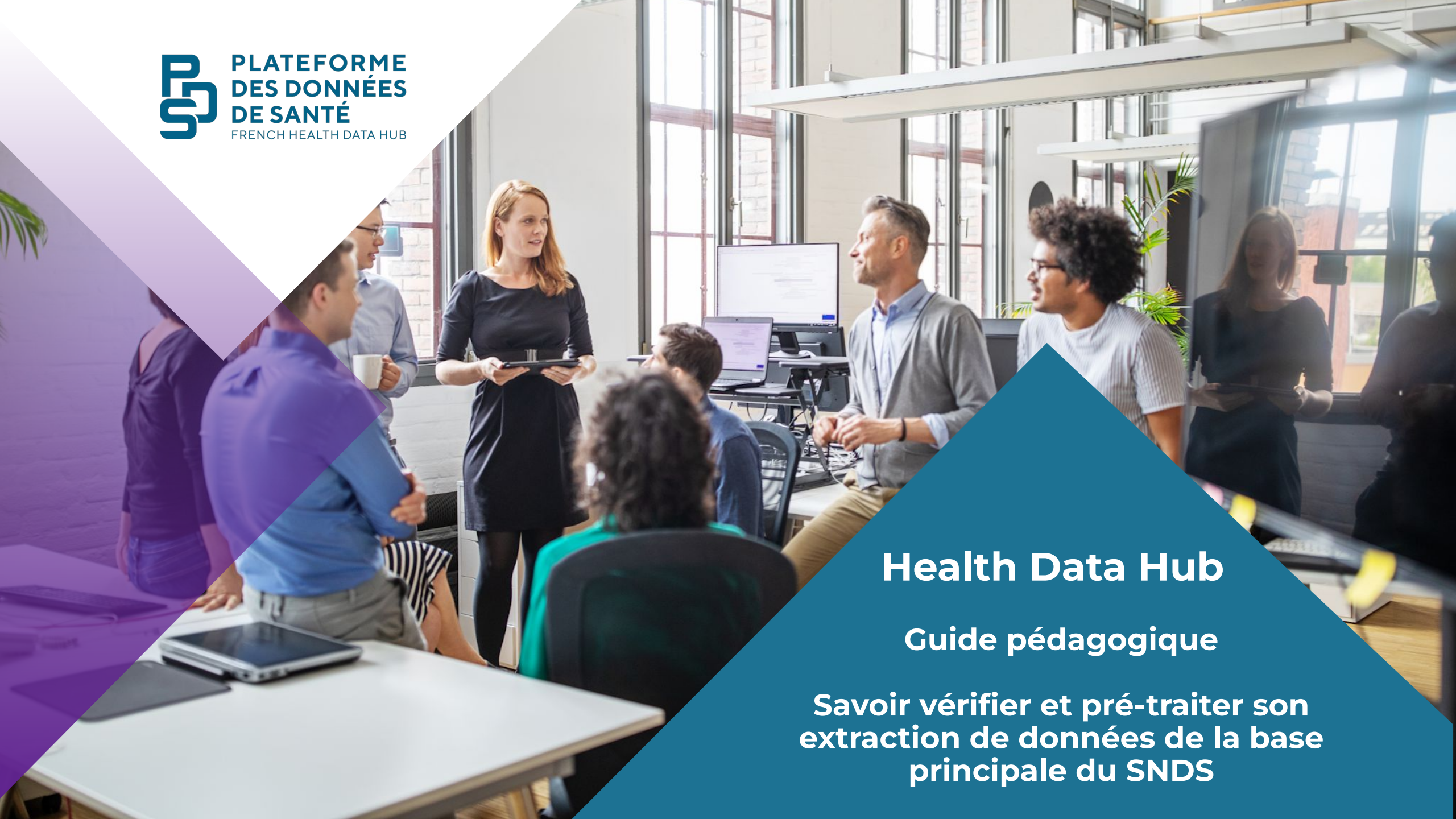
PLATEFORME DES DONNÉES DE SANTÉ
FRENCH HEALTH DATA HUB



Health Data Hub

Guide pédagogique

Savoir vérifier et pré-traiter son extraction de données de la base principale du SNDS

Nos remerciements

Ce guide pédagogique a été réalisé en partenariat avec la Direction du Numérique (DNUM) du Ministère de la Santé, des Familles, de l'Autonomie et des Personnes Handicapées et a bénéficié de l'expertise de **Jérôme Brocca**, Chef de projet Données de santé, en particulier pour la construction du jeu de données et du cas pratique.

Nous tenons également à remercier **Laure Fernandez de Martini** et **Sophie Fantin** du Pôle Accompagnement du SNDS de la Caisse Nationale d'Assurance Maladie pour la qualité de leurs contributions constructives, leurs retours et leurs relectures attentives qui ont permis d'enrichir le contenu de ce guide.

Sommaire de la formation

1. Contexte introductif
   - Généralités sur la nécessité de pré-traiter ses données

2. Vérifier son extraction de données de la base principale du SNDS
   - Étapes génériques pour vérifier son extraction de données

3. Pré-traiter son extraction de données de la base principale du SNDS
   - Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes
   - Gestion des doublons entre DCIR et PMSI : caractéristiques des bénéficiaires
   - Gestion des doublons entre DCIR, PMSI et CépiDC : date de décès
   - Gestion des doublons entre DCIR, PMSI et CépiDC : actes médicaux
   - Supprimer les informations erronées au sein de son extraction
   - Récupération des identifiants uniques
   - Glossaire des noms des variables permettant de qualifier les bénéficiaires dans la base principale du SNDS
   - Appliquer des filtres dans son extraction (données du PMSI)

4. Exercices pratiques



Références

Cette formation se base notamment sur :



- Les ressources documentaires de la CNAM, partenaire institutionnel ;
- Le Module 4 du MOOC SNDS *"Les causes médicales de décès, autres concepts du SNDS et règles de data management"* ;
- Les ressources pédagogiques du site sur la documentation collaborative du SNDS ;
- Le *"Guide d'utilisation du Système National des Données de Santé pour la surveillance et l'observation épidémiologiques (juillet 2024)"*, de Santé Publique France (SpF) ;
- La présentation « Help SNDS » sur : "Le périmètre des données (PMSI / DCIR) - exemple concret autour des actes CCAM", de la Direction du Numérique et de l'Agence Régionale de Santé (ARS) ;
- Le programme "Jointure d'une table avec les référentiels bénéficiaires (IR_BEN_R et IR_BEN_R_ARC)" développé par SpF et mis à disposition dans la Bibliothèque ouverte d'algorithmes en santé (BOAS) du Health data Hub (HDH)
- La géographie des données du bénéficiaire dans le SNDS, de la Direction du Numérique.

Présentation de l'objet du module

- L'objet de ce module est **d'introduire** les utilisateurs de la base principale* (BP) du Système National des Données de Santé (SNDS) **aux règles de vérification et pré-traitement des données afin d'en supprimer les artefacts et anomalies rencontrés**.

- Ce module se divise en 2 volets :
  - **un guide** qui présente **la partie théorique** de ce module incluant des programmes SAS dans un document en annexe, permettant de réaliser certaines étapes présentées dans les diapositives
  - **un exercice pratique** (disponible via ce lien) à réaliser de façon autonome et qui permettra aux utilisateurs de s'exercer à manipuler une base de données synthétiques de la BP du SNDS, **restitué au format dans le cadre d'un accès permanent****, comprenant des données fictives du Datamart de Consommation Inter-Régimes (DCIR) et du Programme de Médicalisation du Système d'Information (PMSI) reproduisant fidèlement la structure et le réalisme des données.

Pour des données extraites **dans le cadre d'un accès sur projet**, c'est la CNAM qui réalise l'extraction des données et donc le ciblage.

* La base principale du SNDS comprend les données du SNDS historique (données du DCIR, du PMSI, du CépiDC, données de handicap de la CNSA) et les données des bases exhaustives appariées à la base historique (données de dépistage COVID-19 SIDEP et de vaccination COVID-10 VAC-SI).

** Les noms des variables présentés dans les exercices illustrés et les exercices pratiques sont ceux figurant dans les données en accès permanent. Certaines différences peuvent exister avec le nom des variables dans le cadre d'un accès sur projet.



1. Contexte introductif

Généralités sur la nécessité de pré-traiter ses données (1/2)

La **base principale du SNDS (Système National des Données de Santé)** a été conçue pour **le remboursement des soins et la gestion des dépenses de santé**. Sa structure et le type d'information qu'elle contient sont principalement **optimisés pour des finalités comptables et administratives**. Plusieurs types d'artefacts ou d'anomalies peuvent être rencontrés dans les données :

Doublons ou redondances d'informations

Certaines données peuvent être présentes dans plusieurs domaines (DCIR, PMSI, CépiDc)

Informations erronées

Présence de données incohérentes, telles que date ou âge aberrants, nécessitant un filtrage, un recodage ou une suppression selon des règles métier adaptées

Discordances sur les dates de décès

Ces informations peuvent se trouver dans différentes tables de différents domaines avec des valeurs divergentes

Incohérences intra-tables

Anomalies internes aux tables de données qu'il convient d'identifier puis, le cas échéant, de corriger ou d'exclure

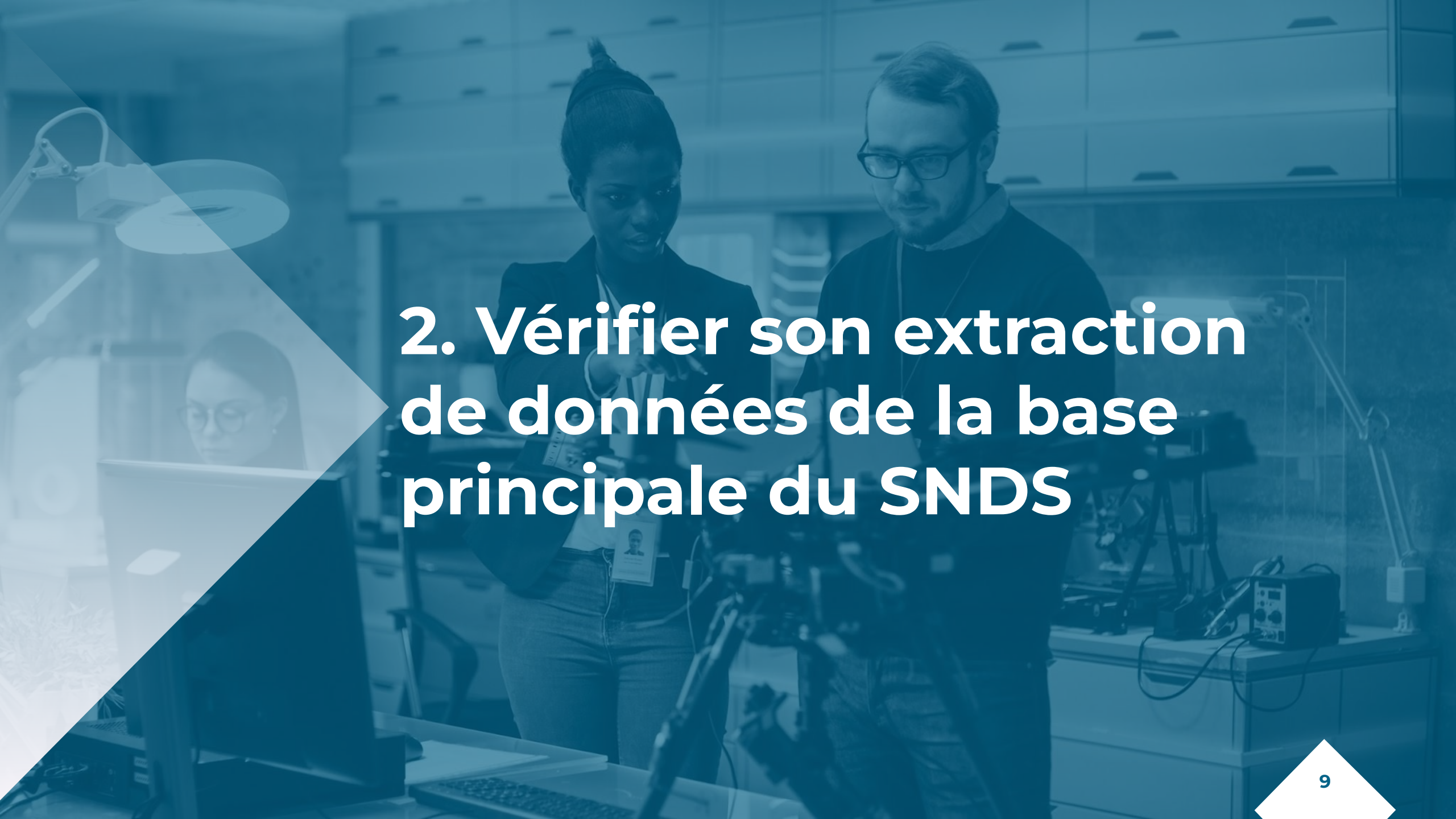
Généralités sur la nécessité de pré-traiter ses données (2/2)

Un **pré-traitement des données** est donc conseillé et peut être réalisé de façon automatique à chaque extraction de données selon les besoins et objectifs du projet. Ce pré-traitement des données vise à :

Fiabiliser les résultats produits à partir de la base principale (BP) du SNDS

Réduire les doublons d'information susceptibles, par exemple, de générer un surcomptage d'actes ou soins
(Remarque : le dénombrement des actes fera l'objet d'un module spécifique)

Garantir la robustesse des analyses



2. Vérifier son extraction de données de la base principale du SNDS

Étapes génériques pour vérifier son extraction de données (1/2)

Avant de traiter les données qui ont été extraites de la base principale du SNDS (par l'utilisateur lui-même dans le cas d'un accès permanent ou par la CNAM dans le cas d'un accès sur projet), il est important de s'assurer que **la sélection des patients est correcte**, que **les fichiers extraits sont les bons**, que **la période extraite correspond bien à l'attendu** et que **les critères figurant dans le protocole d'étude ou dans la fiche d'expression des besoins (EDB - accès sur projet) ont été suivis.**

Dans le cas d'une extraction via un accès sur projet :

1. Contrôler **la cohérence de la livraison**
Comparer le nombre de tables/lignes/colonnes mentionnées dans le fichier Excel de Bilan de livraison fourni par la CNAM (bilan d'extraction) avec le nombre de tables/lignes/colonnes effectivement présentes dans l'extraction reçue

2. Dénombrer **les NUM_ENQ et NUM_ENQ_ANO de la table centrale IR_BEN_R** afin d'estimer approximativement le nombre de bénéficiaires du SNDS présents dans la cohorte livrée et comparer avec les effectifs attendus et indiqués dans le Bilan de livraison fourni par la CNAM

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

Étapes génériques pour vérifier son extraction de données (2/2)

Dans le cas d'une extraction via un accès sur projet (suite) :

3 Comparer **les données extraites** (ou livrées) **avec ce qui est autorisé par la CNIL**
S'assurer de la présence des tables des composants requis (PMSI MCO, DCIR, CépiDC, etc..), de la temporalité des données (période d'extraction souhaitée) et des tables de nomenclatures

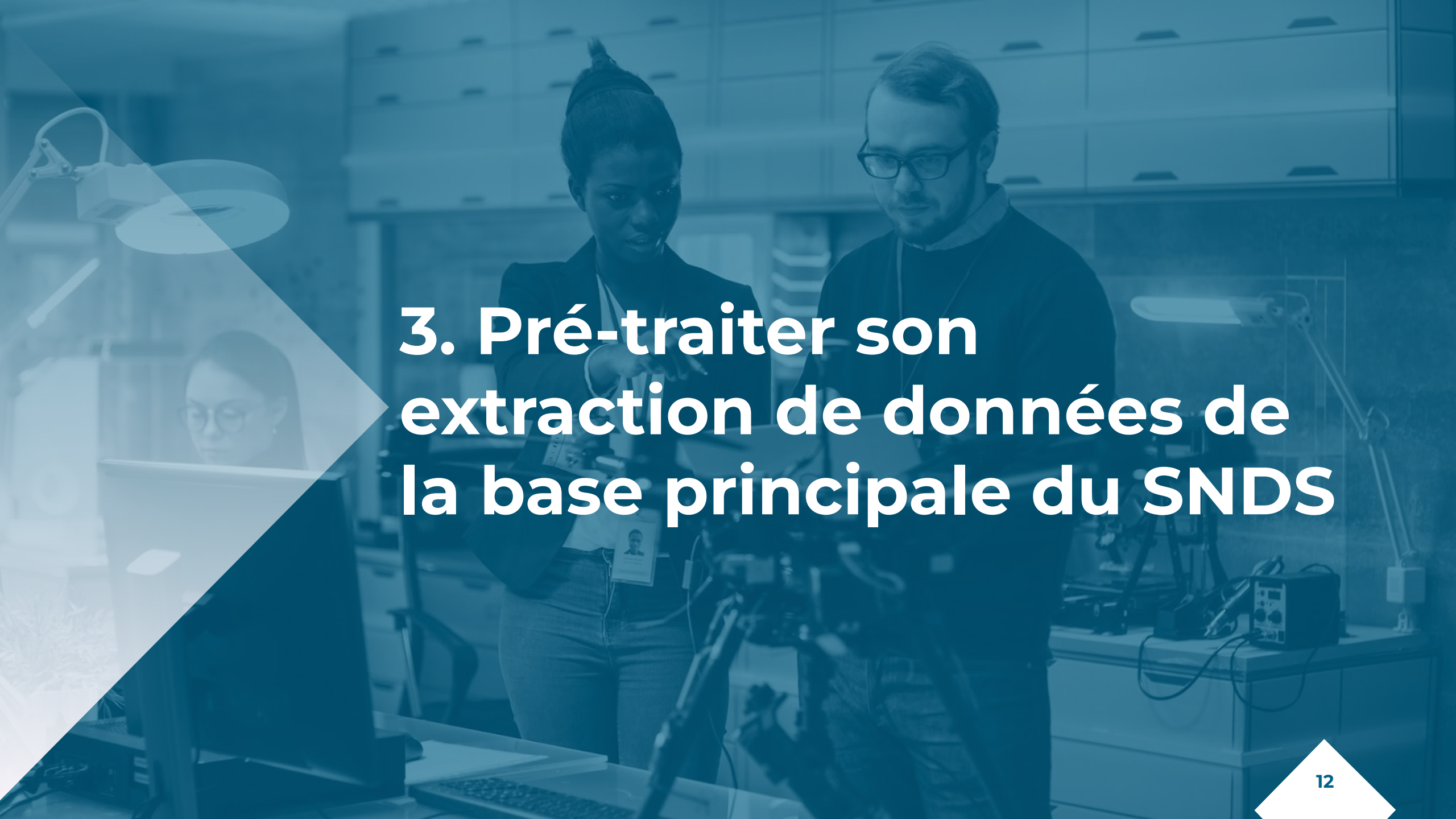
4 Lister et s'assurer que **les variables sensibles autorisées** (Ex : date de soins complètes, code de la commune de résidence, mois de naissance, date de décès complète, code de la commune de décès...) **soient présentes**, et que **les variables sensibles non autorisées soient absentes. Dans le cas contraire, il faudra les supprimer des données extraites.**

Dans tous les cas de figure (accès permanent ou sur projet) :

5 Vérifier **la présence et le bon format des variables d'identifiants pseudonymisés** (NUM_ENQ / BEN_NIR_PSA / NIR_ANO_17, BEN_IDT_ANO / NUM_ENQ_IDT et BEN_NIR_ANO / NUM_ENQ_ANO) et **du rang gémellaire** (BEN_RNG_GEM) **dans les tables de référentiel des bénéficiaires** (IR_BEN_R/IR_BEN_R_ARC) **ainsi que dans les autres tables censées les contenir** *(Ex : ER_PRS_F, IR_IMB_R, T_MCOxxC...)* pour permettre des jointures correctes entre ces tables.

6 Vérifier **la quantité mensuelle de soins et de sujets dans chaque table** (ER_PRS_F, tables affinées de DCIR, T_MCOxxB, T_SSRxxB...) afin de contrôler si de fortes hausses ou baisses de quantité de soins sont observées dans les données extraites et ainsi identifier tout problème d'extraction ou de chargement des données. Des tableaux et graphiques réalisés sur la fréquence mensuelle des soins ou des sujets distincts peuvent être édités pour cela (voir diapositive suivante)

Des programmes SAS illustrant ces éléments sont disponibles en annexe.



3. Pré-traiter son extraction de données de la base principale du SNDS

Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes (1/5)

Des **séjours ou données hospitalières** peuvent se retrouver en double dans DCIR et le PMSI.

Le pré-traitement de son extraction permet alors de choisir entre les données issues du DCIR et les données issues du PMSI les plus pertinentes au regard des besoins de son étude.

DCIR : remboursements de soins

PMSI : financement des hôpitaux

DCIR	PMSI
Activité en cabinets de ville, en centres de santé (table ER_PRS_F)	**Non**
Séjours en établissements privés : données financières et non médicales (tables ER_PRS_F et ER_ETE_F, *si PRS_PPU_SEC = 2)*	**Séjours en établissements privés : données médicales** (tables TMCOxxB et T_MCOxxE, *si STA_ETA = 'OQN'*)
Non *(à l'exception d'1 seul centre hospitalier en facturation FIDES : l'Hopital ESPIC Wallerstein)*	**Séjours en établissements publics** (tables TMCOxxB et T_MCOxxE, *si STA_ETA = 'STC'*)
Activité (ACE, séjours) des établissements publics transmise pour information (pas de facturation directe) (table ER_PRS_F) **Activité en Actes et consultations externes (ACE) des établissements publics en facturation directe (FIDES)** (table ER_PRS_F)	**Activité en ACE des établissements publics exhaustive** (tables T_MCOxxFASTC et T_MCOxxVALOACE)

Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes (2/5)

Cas des établissements du secteur public ex-Dotation Globale (DG)

Cela concerne les établissements publics de santé et les établissements privés à but non lucratif participant au service public hospitalier

Les **remontées d'information sur les prestations** réalisées dans les ES ex-DG sont **très incomplètes dans le DCIR et en doublon avec le PMSI**. C'est pourquoi, il est **conseillé d'exclure l'information concernant les hôpitaux publics** (en facturation FIDES ou non FIDES) **dans le DCIR**, afin de privilégier leur étude via le PMSI.

Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes (3/5)

Cas des établissements du secteur privé ex-Objectif Quantifié National (OQN)

Pour les données d'activité des séjours hospitaliers des ES ex-OQN (secteur privé), le choix de la source de données va dépendre de la finalité de l'étude

Etudes épidémiologiques :

Les données de séjours du PMSI (privés et publics) contiennent des informations médicalisées, et particulièrement les diagnostics du séjour qui seront utiles à ce type d'études notamment pour calculer les taux d'incidence / prévalence de pathologies.

Etudes économiques (suivi des dépenses ...) :

Les séjours en établissements privés sont facturés directement à l'AM ce qui garantit l'exhaustivité des remontées d'information sur ce champ. Cela concerne toutes les prestations en établissement privé, que l'établissement soit de nature non lucratif ou lucratif.

De ce fait, les données du DCIR contiennent les montants payés et remboursés de chaque séjour en établissement privé. Le PMSI contient également des informations de facturation des séjours privés mais il n'est pas certain qu'elles correspondent aux montants réellement facturés et remboursés par l'Assurance Maladie.

Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes (4/5)

L'information relative à **l'activité en actes et consultations externes (ACE)** est remontée depuis 2009 dans la table des prestations ER_PRS_F (DCIR), elle est disponible :

- Dans la variable **DPN_QLF** qui contient le qualificatif de la dépense de soin, permettant de distinguer notamment les dépassements d'un professionnel de santé appliqués sur une prestation (permanent, maîtrisé, autorisé, justifié…).
- Dans une 2ème variable **PRS_DPN_QLP** de qualificatif de la dépense (transmis PS5), renseignée dans le cadre du dispositif des participations forfaitaires assurés.

- Ces **soins externes ACE, qui ne donnent pas lieu à une facturation à l'activité,** sont transmis "pour information" à l'Assurance Maladie **via le code '71'** mais **ne sont pas exhaustifs dans DCIR** (et de qualité inconnue).
- En revanche, **ils sont exhaustifs dans le PMSI** (Exemple : table T_MCOxxVALOACE) car ils relèvent du financement par budget global.

Gestion des doublons entre DCIR et PMSI : données de séjours, actes et consultations externes (5/5)

2 **L'activité externe des établissements publics facturant à l'activité** (TAA ou T2A) via l'assurance maladie (facturation FIDES) est également présente dans la table ER_PRS_F. Pour identifier ces soins, il faut relier ER_PRS_F avec ER_ETE_F (table du détail des informations liées à l'exécution de la prestation dans un établissement) qui contient :

- La variable **ETE_IND_TAA** (indicateur T2A) dont le **code '1'** correspond à l'activité T2A publique



Il est recommandé d'exclure ces soins

1	Par la condition *DPN_QLF = 71*, on identifie **l'activité externe des établissements publics en budget global** transmises à l'assurance maladie pour information.
2	Par la condition *DPN_QLF = 0 AND PRS_DPN_QLP = 71*, on identifie **les participations forfaitaires assurés payées pour les consultations externes**, transmises à l'Assurance Maladie pour connaître l'ensemble des participations forfaitaires des assurés.
3	Par la condition *ETE_IND_TAA = 1*, on identifie l'**activité externe des établissements publics facturant à l'activité (TAA)**, transmise à l'Assurance Maladie *(Note : ETE_IND_TAA = NULL lorsque les soins sont réalisés en cabinet libéral)*



Synthèse : Pour exclure des analyses toutes les consultations externes et forfaits transmis "pour information", y compris les participations forfaitaires qu'elles ont générées, **ainsi que l'activité externe des hôpitaux publics pratiquant la facturation directe**, on peut utiliser le filtre suivant (après avoir relié avec un LEFT JOIN les tables ER_PRS_F et ER_ETE_F sur les 9 clés de jointure) :

Ex : WHERE DPN_QLF NOT IN (71) AND PRS_DPN_QLP NOT IN (71) AND (ETE_IND_TAA <> 1 OR ETE_IND_TAA IS NULL)

Gestion des doublons entre DCIR et PMSI : caractéristiques des bénéficiaires (1/2)

Outre la date de décès, plusieurs informations relatives à l'individu sont également redondantes dans les tables IR_BEN_R et ER_PRS_F (DCIR) et dans le PMSI, telles que :

	DCIR	PMSI			
Tables source	*(Tables IR_BEN_R, ER_PRS_F)*	**MCO** *(Tables T_MCOxxB, T_MCOxxFA, T_MCOxxFASTC)*	**HAD** *(Tables T_HADxxB, T_HADxxFA)*	**SSR** *(Tables T_SSRxxB, T_SSRxxFA, T_SSRxxFASTC)*	**RIM-P** *(Tables T_RIPxxR3A, T_RIPxxRSA, T_RIPxxFA)*
Sexe	BEN_SEX_COD	COD_SEX	COD_SEX	COD_SEX	COD_SEX
Lieu de résidence du bénéficiaire : - **commune** - **département**	BEN_RES_COM BEN_RES_DPT	BDI_COD BDI_DEP	BDI_COD BDI_DEP	BDI_COD BDI_DEP	BDI_COD BDI_DEP
- **code postal**	X	COD_POST, POST_DEP	COD_POST, POST_DEP	COD_POST, POST_DEP	COD_POST, POST_DEP



Concernant le lieu de résidence des bénéficiaires :

- Pour le DCIR, la source du code commune est l'INSEE
- Pour le PMSI, c'est le code géographique du PMSI qui correspond plus ou moins au code postal (voir le site de l'ATIH)

Gestion des doublons entre DCIR et PMSI : caractéristiques des bénéficiaires (2/2)

- La table IR_BEN_R contient ces variables avec la **dernière information connue** par l'Assurance Maladie sur le couple BEN_NIR_PSA||BEN_RNG_GEM
- La table ER_PRS_F et celles du PMSI contiennent ces informations **au moment de l'exécution du soin.**

Selon la finalité de l'étude, il peut être recommandé d'utiliser seulement les informations présentes dans le référentiel bénéficiaire IR_BEN_R car il s'agit de données administratives collectées par l'Assurance Maladie auprès des bénéficiaires.

Vous trouverez plus de détails sur les informations relatives à la géographie du bénéficiaire dans la base principale du SNDS dans la présentation de la Direction du Numérique (DNUM) suivante : https://view.genially.com/67c93b8930fe19f52e051dab

Gestion des doublons entre DCIR, PMSI et CépiDc : date de décès (1/3)

Les informations relatives aux dates de décès des individus peuvent se situer dans différentes tables. La date de décès peut être présente de façon directe ou par croisement entre deux variables.

Le tableau ci-dessous recense ces informations :

Source	Table(s)	Variable(s)
DCIR	ER_PRS_F	BEN_DCD_DTE / BEN_DCD_AME
DCIR	ER_PRS_F + ER_ETE_F	EXE_SOI_DTD (si ETE_NAT_FSJ='D')
Référentiel bénéficiaire	IR_BEN_R +/- IR_BEN_R_ARC	BEN_DCD_DTE / BEN_DCD_AME
PMSI MCO	T_MCOaaC + T_MCOaaB	SOR_DAT (si SOR_MOD=9)
PMSI SSR	T_SSRaaC + T_SSRaaB	SOR_DAT (si SOR_MOD=9)
PMSI HAD	T_SSRaaC + T_SSRaaB	SOR_DAT (si SOR_MOD=9)
PMSI PSY	T_SSRaaC + T_SSRaaRSA	SOR_DAT (si SOR_MOD=9)
CépiDc	KI_CCI_R	BEN_DCD_DTE / BEN_DCD_AME

Gestion des doublons entre DCIR, PMSI et CépiDc : date de décès (2/3)

L'alimentation et les spécificités de la date de décès selon la source de données considérée sont détaillées ci-dessous.

Tables	Informations
ER_PRS_F	Les dates de décès ne remontent que pour les individus dont une prestation a fait l'objet d'un remboursement après le décès (Ex : lors du versement du capital décès - pour les salariés - ou du versement de prestations accompagnant la fin de vie) et pour qui la date de décès est renseignée.
IR_BEN_R	Les dates de décès sont alimentées par ER_PRS_F et par la Base de données opérante (BDO) dont le flux provient du référentiel individu de Versailles qui obtient l'information des certifications de décès par l'INSEE. Les décès concernés sont ceux survenus en France et à l'étranger. Remarque : La date de décès n'est remontée que pour le dernier BEN_NIR_PSA / NUM_ENQ. Elle n'est donc pas renseignée dans IR_BEN_R pour l'ensemble des BEN_NIR_PSA / NUM_ENQ associés au même bénéficiaire : un bénéficiaire peut être vivant (avec un BEN_NIR_PSA lors de son enfance) et mort (avec son BEN_NIR_PSA adulte).
PMSI	Il peut arriver que des codes de mode de sortie de séjour '9' (décès) soient attribués par erreur. La date de décès d'un patient non affilié au Régime général (RG) et décédé à l'hôpital (ou dont la date n'est pas remontée dans la base de données opérante) et qui n'a jamais consommé de soins de ville ne remonte pas dans IR_BEN_R. C'est le cas notamment d'une partie des décès néonataux ou des patients ayant plusieurs pseudo-NIR.
KI_CCI_R	Les dates de décès sont obtenues à partir des certificats de décès et sont alignées avec celles de l'INSEE, pour les décès survenus en France uniquement (à la différence d'IR_BEN_R qui intègre aussi les décès survenus à l'étranger). L'appariement indirect de la date et des causes médicales de décès du CépiDC avec IR_BEN_R ne couvre pas l'ensemble des individus du SNDS (le taux d'appariement était de 87% en 2022) et il y a aussi un décalage temporel dans le chargement de ces données au sein du SNDS (au 01/01/2026, l'année 2023 constitue la plus récente année de décès disponible). **L'ensemble de ces éléments peut expliquer des divergences de date de décès parfois observées entre IR_BEN_R et KI_CCI_R.**

Gestion des doublons entre DCIR, PMSI et CépiDc : date de décès (3/3)

Différentes méthodes peuvent être envisagées pour déterminer la date de décès en cas d'informations discordantes. Pour chacune d'elles, il est conseillé de recenser l'information aussi largement que possible depuis les différentes tables contenant l'information de date de décès. La seconde étape consiste alors à choisir quelle information prioriser et conserver lorsque celle-ci est présente dans plusieurs tables et qu'elle n'est pas uniforme dans toutes les sources.

Pour cela, voici deux exemples de règles à appliquer (règles non obligatoires) :

1

Choisir la date de décès la plus récente

2

Etablir un algorithme donnant priorité à l'une ou l'autre des sources

(Ex : présence de dates de décès discordantes dans KI_CCI_R et IR_BEN_R ⇒ choix de privilégier celle d'IR_BEN_R)



Il peut arriver que **la date de décès ne soit pas renseignée dans la table des bénéficiaires IR_BEN_R** alors qu'une **date de décès** est indiquée dans la table **KI_CCI_R du CépiDC (ou inversement)** ou qu'un **décès** a été enregistré lors d'une **hospitalisation MCO (T_MCOxxB, séjour avec mode de sortie '9')**.

➔ Il est essentiel de **croiser ces différentes sources, de vérifier la présence ou non de soins après les dates de décès discordantes** afin d'affiner sa décision sur le choix de la date de décès et d'éviter des analyses reposant sur des données de mortalité erronées, notamment s'il s'agit d'un critère important de votre étude.

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

Gestion des doublons entre DCIR et PMSI : actes médicaux (1/2)

Des informations sur les **actes et procédures médicales** sont disponibles dans DCIR et le PMSI et certaines sont en double.

Vous trouverez plus de détails sur ces éléments (notamment des programmes associés) dans la présentation de l'ARS « Help SNDS » sur : "Le périmètre des données (PMSI / DCIR) - exemple concret autour des actes CCAM".

DCIR	PMSI
Actes CCAM en cabinets de ville et centres de santé *(table ER_CAM_F)*	**Non**
Actes CCAM au cours de séjours en établissements privés *(tables ER_PRS_F, ER_CAM_F et ER_ETE_F)*	**Actes CCAM au cours de séjours en établissements privés** *(tables T_MCOxxB, T_MCOxxA et T_MCOxxE, si STA_ETA = 'OQN')*

Il est recommandé d'utiliser les données de DCIR pour les actes médicaux en établissements de soins privés (plus exhaustifs, tarifs…)

Non	**Actes CCAM au cours de séjours en établissements publics** *(tables T_MCOxxB, T_MCOxxA et T_MCOxxE, si STA_ETA = 'DGF')*
Actes CCAM de l'activité en ACE des établissements publics transmise pour information *(tables ER_PRS_F et ER_CAM_F)* **ou en facturation directe FIDES** *(tables ER_PRS_F, ER_CAM_F)*	**Actes CCAM de l'activité en ACE en établissements publics** *(tables T_MCOxxCSTC et T_MCOxxFMSTC)*

Il est recommandé d'exclure ces soins *(cf. diapositive 19)*

Gestion des doublons entre DCIR et PMSI : médicaments (2/2)

Des informations sur les **médicaments hospitaliers** sont disponibles dans DCIR et le PMSI :

- **médicaments de la liste en sus** : facturés à l'Assurance Maladie en sus de la tarification du séjour (GHS) car onéreux ou innovants ;
- **médicaments rétrocédés :** certains établissements de santé disposant d'une pharmacie à usage interne (PUI) peuvent être autorisés à dispenser des médicaments à des patients non hospitalisés car ils présentent des contraintes particulières de distribution, de dispensation ou d'administration, ou nécessitent un suivi de la prescription ou de la délivrance.

DCIR	PMSI
Médicaments délivrés en officine de ville *(table ER_PHA_F)*	**Non**
Médicaments rétrocédés des établissements publics (table ER_UCD_F, *si UCD_TOP_UCD = 0*)	**Non**
Médicaments rétrocédés des établissements privés (table ER_UCD_F, *si UCD_TOP_UCD = 0*)	**Non**
Non	**Médicaments en sus du GHS des établissements publics** *(table T_MCOxxMED, T_MCOxxFHSTC)*
Médicaments en sus du GHS des établissements privés (table ER_UCD_F, *si UCD_TOP_UCD = 1*)	**Médicaments en sus du GHS des établissements privés** *(table T_MCOxxFH)*



Pour les médicaments en sus du GHS des établissements privés, il est conseillé de **privilégier la table ER_UCD_F** car celle-ci contient la tarification des ces médicaments alors que c'est plutôt informatif dans le PMSI.

Supprimer les informations erronées au sein de son extraction (1/4)

Des informations erronées dans les données de la base principale du SNDS peuvent également se retrouver sous la forme suivante *(liste non exhaustive)* :

- Un soin remboursé ou exécuté postérieurement à la date de décès
  Ex : IF EXE_SOI_DTD > BEN_DCD_DTE THEN DELETE
- Un soin dont la date de réalisation (EXE_SOI_DTD) est manquante
  Ex : IF EXE_SOI_DTD = . THEN DELETE
- Un soin de la table ER_PRS_F dont la nature de prestation est “Sans objet”
  Ex : IF (PRS_NAT_REF = 0) THEN DELETE
- Un patient avec une année de naissance incohérente
  Ex : WHERE BEN_NAI_ANN > 1900
- Un acte d’accouchement chez un bénéficiaire de sexe masculin, une chirurgie de la prostate chez un bénéficiaire de sexe féminin…
- …



Ces erreurs peuvent être liées à de mauvaises informations renseignées par le professionnel de santé ou l’établissement, ou à des anomalies de transmission dans les flux de données (= artefact de codage ou de transmission).

Remarque : de manière générale, ***ces données erronées sont présentes en faible quantité*** *dans la base principale du SNDS.*

Le **pré-traitement des données extraites** permet, entre autres, de filtrer les incohérences, de recoder ou supprimer les valeurs extrêmes et d’exclure certaines données selon des règles métier et selon ses besoins.

Supprimer les informations erronées au sein de son extraction (2/4)

Lorsqu'on utilise le référentiel bénéficiaire IR_BEN_R, il est conseillé de sélectionner **des identifiants certifiés par l'INSEE**, et éventuellement des identifiants provisoires (migrant provisoire ou ouvrant de droit provisoire). La variable **BEN_CDI_NIR** présente dans les tables IR_BEN_R et ER_PRS_F est le code d'identification du NIR et permet de distinguer :

1- Identifiants certifiés par l'INSEE (NIR normal) : **BEN_CDI_NIR = 00**
Le NIR bénéficiaire est certifié lorsque l'identité de la personne (noms, prénoms, date et lieu de naissance) a été validée par l'INSEE. La condition BEN_CDI_NIR = 00 pour sélectionner les individus de sa population d'étude est un filtre qui permet de supprimer les bénéficiaires fictifs et provisoires.

2- Identifiants provisoires : **BEN_CDI_NIR = 03 ou 04**
Exemple : les bénéficiaires de l'Aide médicale d'état (AME) n'ont pas de carte vitale et se voient attribuer un numéro de sécurité sociale temporaire ; le NIR attribué par un régime à un travailleur ou un étudiant étranger dans l'attente de la validation des documents permettant de confirmer son identité.

3- Identifiants fictifs : **BEN_CDI_NIR = autres valeurs de codage**
Des NIR fictifs existent pour certaines prestations (Ex : IVG, campagne de vaccination scolaire papillomavirus, COVID-19 etc.) afin de garantir l'anonymat de la personne (pour ses autres prestations le bénéficiaire garde son identifiant habituel), ou également pour des rémunérations sur objectif des professionnels de santé (avec aucun « vrai » patient associé). Lorsque le NIR est fictif :

- L'identifiant bénéficiaire BEN_NIR_ANO n'est pas renseigné ;
- Dans ER_PRS_F, le code de petit régime d'affiliation est RGM_COD = 888 avec un grand régime de liquidation RGM_GRG_COD = 01 ;
- Leurs données socio-démographiques sont la plupart du temps erronées.



Dans le cadre d'une étude du parcours de soins individuel, il est conseillé de conserver les individus pour lesquels le NIR est certifié
Ex : WHERE BEN_CDI_NIR = 00

Supprimer les informations erronées au sein de son extraction (3/4)

La **variable BEN_AMA_COD** dans la table ER_PRS_F de DCIR contient **l'âge du bénéficiaire à la date du soin**. Cette variable est calculée comme le délai entre l'année et le mois de la date du soin EXE_SOI_DTD (ou à défaut de la date de liquidation) et l'année et le mois de naissance du bénéficiaire. Elle est exprimée en mois / années révolu(e)s :

➔ **bénéficiaires < 2 ans : BEN_AMA_COD = 1 000 + âge en mois MM** (MM = [00-23 mois], c'est à dire l'âge en mois révolus)
Exemple : Un nourrisson a la valeur BEN_AMA_COD = 1001 jusqu'à ses 2 mois, etc. jusqu'à 1023 quand il a 23 mois. Ensuite, il a la valeur BEN_AMA_COD = 2 à partir du jour de son 2ème anniversaire.

➔ **bénéficiaires ≥ 2 ans : BEN_AMA_COD = XXX** (XXX représente l'âge en années révolues)
Exemple : une personne a la valeur BEN_AMA_COD = 25 le jour de son 25ème anniversaire, et aura la valeur BEN_AMA_COD = 26 lors de son 26ème anniversaire.

Nota bene

- ***Attention*** *: il peut parfois arriver de rencontrer le cas où BEN_AMA_COD = 1024 lorsque le sujet est âgé de 2 ans*
- *Lorsque l'âge calculé est > 129 ou < 0, BEN_AMA_COD prend la valeur 9999*
- *Si des patients âgés sont ciblés, alors il faut mettre des bornes sur la variable BEN_AMA_COD pour éviter de compter à tort des nourrissons.*

Il est préconisé de pré-traiter ses données en ne conservant que les âges et sexe corrects. Ici, un exemple chez des sujets dont le NIR est certifié :
Exemple : WHERE (002≤BEN_AMA_COD≤110 OR 1000≤BEN_AMA_COD≤1024) AND BEN_SEX_COD in (1,2) AND BEN_CDI_NIR=00

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

Supprimer les informations erronées au sein de son extraction (4/4)

Dans le référentiel des bénéficiaires IR_BEN_R :

1. **L'année de naissance du bénéficiaire (variable BEN_NAI_ANN)** sert notamment à calculer l'âge. Sa valeur est forcée à '1600' lorsqu'elle est inconnue et les mois-année de naissance entièrement manquants sont codés '01' et '1600'. Pour éviter de calculer des âges aberrants, il convient de remplacer ces valeurs fictives par des valeurs vides avant tout calcul.

   Ex : IF BEN_NAI_ANN = 1600 THEN BEN_NAI_ANN = . ;

2. **La date de décès (BEN_DCD_DTE)** suit la même logique : elle est renseignée avec la valeur '01/01/1600' lorsqu'aucun décès n'est enregistré (bénéficiaire vivant ou date de décès inconnue). Il est donc fortement recommandé de remplacer cette valeur fictive par une valeur vide afin d'éviter des calculs erronés (exemple : âge au décès, délai de survie depuis une date de référence). C'est également le cas pour la variable BEN_DCD_AME (année-mois de décès de type AAAAMM) mis à '160001' en cas d'absence d'information de décès.

   Ex : IF YEAR(BEN_DCD_DTE) = 1600 THEN BEN_DCD_DTE= . ;

3. **La date d'insertion (BEN_DTE_INS)** prend également la valeur '01/01/1600' lorsqu'elle n'est pas renseignée.

   Ex : IF YEAR(BEN_DTE_INS) = 1600 THEN BEN_DTE_INS= . ;

Récupération des identifiants uniques



Dans le cadre du ciblage d'un groupe de bénéficiaires répondant à des critères de sélection définis, on récupère en 1er lieu leur pseudo-NIR BEN_NIR_PSA et leur rang de naissance BEN_RNG_GEM associé en lien avec ces critères. Il est important de récupérer tous les autres pseudo-NIR et rang de naissance de ces individus afin de pouvoir reconstituer la totalité de leur consommation de soins sur la période d'étude considérée, y compris sur les périodes antérieures ou postérieures au ciblage initial.

1

Étape 1 : Ciblage initial

Critères de sélection : pathologie CIM-10, acte CCAM, médicament CIP…

↓

Sélection de bénéficiaires (DCIR et/ou PMSI)

↓

Liste d'identifiants NIR_ANO_17 *(si issus du PMSI)* et de **couples BEN_NIR_PSA||BEN_RNG_GEM** *(si issus de DCIR)* **liés aux critères**

2

Étape 2 : Rattachement identifiant unique

Relier **NIR_ANO_17** *(issu du PMSI)* à **BEN_NIR_PSA (et BEN_RNG_GEM)** via IR_BEN_R et IR_BEN_R_ARC *

↓

Puis, pour chaque couple (BEN_NIR_PSA, BEN_RNG_GEM) *(issus de DCIR et PMSI)* **:**

→ Rechercher BEN_IDT_ANO correspondant dans IR_BEN_R et IR_BEN_R_ARC

↓

1 (BEN_NIR_PSA, BEN_RNG_GEM) ↔ 1 BEN_IDT_ANO

** si un NIR_ANO_17 est lié à différentes valeurs de BEN_RNG_GEM avec mêmes sexe et date de naissance, il vaut mieux supprimer ces sujets - **au choix de l'utilisateur***

3

Étape 3 : Extension à tous les pseudo NIR du bénéficiaire

Pour chaque BEN_IDT_ANO distinct retrouvé :

→ Récupérer tous les couples (BEN_NIR_PSA, BEN_RNG_GEM) associés dans IR_BEN_R et IR_BEN_R_ARC

↓

1 BEN_IDT_ANO ↔ 1 à N (BEN_NIR_PSA, BEN_RNG_GEM)

4

Etape 4 : Reconstitution des consommations

Extraire dans DCIR / PMSI l'ensemble de leurs consommations de soins au cours de la période d'étude considérée pour tous les (BEN_NIR_PSA, BEN_RNG_GEM) associés.

→ A partir des tables ER_PRS_F, T_MCOxxB…

→ Permet de reconstituer le parcours complet du bénéficiaire.



Vous trouverez plus de détails sur ces étapes (notamment techniques, avec les programmes associés) dans la présentation de la CNAM « les Clés du SNDS » sur : *"La méthodologie de ciblage de tous les identifiants patients d'une cohorte dans le SNDS : exemple d'une cohorte de patients diabétiques"*.

Un programme SAS développé par Santé Publique France est également disponible **dans la bibliothèque BOAS** via ce **lien** afin de relier les identifiants (BEN_NIR_PSA, BEN_RNG_GEM) d'une table aux référentiels des bénéficiaires IR_BEN_R et IR_BEN_R_ARC.

Glossaire des noms des variables permettant de qualifier les bénéficiaires dans la base principale du SNDS

Nom de variable source (dans le cadre d'un accès permanent)	Tables du SNDS	Définition	Nom de variable restitué (dans le cadre d'une extraction sur projet)
BEN_NIR_PSA	IR_BEN_R, IR_BEN_R_ARC, ER_PRS_F, IR_IMB_R	Pseudo NIR	NUM_ENQ
NIR_ANO_17	T_*XXXaa*C, T_*XXXaa*CSTC, T_*XXXaa*FASTC, T_*XXXaa*FCSTC, T_*XXXaa*FMSTC **	Pseudo NIR	NUM_ENQ
BEN_RNG_GEM*	IR_BEN_R, IR_BEN_R_ARC, ER_PRS_F, IR_IMB_R	Rang gémellaire	BEN_RNG_GEM
BEN_NIR_ANO	IR_BEN_R, IR_BEN_R_ARC, KI_CCI_R, KI_ECD_R	NIR bénéficiaire unique	NUM_ENQ_ANO
BEN_IDT_ANO	IR_BEN_R, IR_BEN_R_ARC, KI_CCI_R, KI_ECD_R	Correspond à BEN_NIR_ANO s'il existe, sinon à BEN_NIR_PSA || BEN_RNG_GEM	*Non restitué, à construire* NUM_ENQ_IDT
ASS_NIR_ANO	IR_BEN_R, IR_BEN_R_ARC	NIR Assuré	*Non restitué*

* Rang gémellaire pour le Régime Général (y compris SLM) OU, pour les autres régimes, le rang pour différencier les bénéficiaires rattachés au même ouvreur de droit.

** Tables du PMSI : *XXX* = MCO, SSR, HAD ou RIP (si disponible) ; *aa* = année de fin du séjour

Appliquer des filtres dans son extraction (données du PMSI) (1/5)

Des **données erronées** ou **en doublons** sont également présentes dans les tables du PMSI et il est nécessaire **d'appliquer des filtres avant de traiter ces données.**

Séjours/séances : il est recommandé **d'appliquer systématiquement plusieurs filtres**

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

MCO

- **Exclure les FINESS géographiques APHP, APHM et HCL** dont les remontées sont en doublons sur leur FINESS juridique avant 2017 inclus (via la table T_MCOxxB) ;
- **Exclure les séjours avec un groupage GHM en erreur, dans la table T_MCOxxB,** correspondant à :
  - Des *séjours avec informations manquantes ou incohérentes* (diagnostics, âge, date de sortie, etc.) ne permettant pas de grouper le séjour dans un GHM et de lui attribuer un tarif ;
  - De *« faux » séjours générés automatiquement* pour des besoins de facturation (dialyse, activité externe, passage aux urgences, forfaits, etc.) des ES ex-OQN ;
- **Exclure les « faux » résumés de sortie anonymes (RSA) générés automatiquement** (dialyses, activité externes de médecins salariés, ATU, FFM, SE…)
- **Exclure les prestations inter établissements (PIE) via la table T_MCOxxB** : transfert temporaire (< 2 jours) d'un patient dans un autre ES pour réaliser une prestation. La même prestation figure en double dans les remontées des 2 ES, mais aucune valorisation ne sera calculée pour l'ES prestataire ;
- **Exclure les séjours avec des clés de chaînage incorrectes** sur les informations des bénéficiaires via les variables codes retours *XXX_RET* de la table T_MCOxxC (***Attention*** *: le nombre de ces variables peut varier selon l'année)*

Appliquer des filtres dans son extraction (données du PMSI) (2/5)

Activité externe : pour travailler sur l'activité externe des établissements de soins ex-DG en MCO, il est recommandé de

MCO

- **Exclure les FINESS géographiques APHP, APHM et HCL** dont les remontées sont en doublons sur leur FINESS juridique avant 2017 inclus via la table T_MCOaaCSTC ;
- **Exclure les séjours avec des clés de chaînage incorrectes** sur les informations des bénéficiaires via les variables codes retours de la table T_MCOaaCSTC.

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

Appliquer des filtres dans son extraction (données du PMSI) (3/5)

Séjours/séquences : Il est recommandé **d'appliquer systématiquement plusieurs filtres**

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

HAD

- **Exclure les sous-séquences** qui ne seront pas valorisées (i.e. les sous-séquences avec une erreur de groupage) via la table T_HADaaGRP : *WHERE GHT_NUM <> '99'* ;
- **Exclure les séjours associés à des clés de chaînage incorrectes sur les informations des bénéficiaires** via les variables codes retours de la table T_HADaaC :
  - Variables disponibles depuis 2005 : *WHERE NIR_RET = '0' AND NAI_RET = '0' AND SEX_RET = '0' AND SEJ_RET = '0' AND FHO_RET = '0' AND PMS_RET = '0' AND DAT_RET = '0'* ;
  - Variables disponibles depuis 2013 : *AND COH_NAI_RET = '0' AND COH_SEX_RET = '0'* ;
- Dans ce champ d'activité, **les ES APHP, APHM et HCL n'ont pas remonté leur activité en double sur leur FINESS géographique.**

Activité externe : pour travailler sur l'activité externe (disponible avec le chaînage des bénéficiaires à partir de 2013), il est recommandé de :

SMR

- **Exclure les ACE associés à des clés de chaînage incorrectes sur les informations des bénéficiaires** via les variables codes retour de la table T_SSRaaCSTC : *WHERE NIR_RET = '0' AND NAI_RET = '0' AND SEX_RET = '0' AND IAS_RET = '0' AND ENT_DAT_RET = '0'* ;
- Dans ce champ d'activité, **les ES APHP, APHM et HCL n'ont pas remonté leur activité en double sur leur FINESS géographique.**

Appliquer des filtres dans son extraction (données du PMSI) (4/5)

Séjours : il est recommandé **d'appliquer systématiquement plusieurs filtres**

Des programmes SAS illustrant ces éléments sont disponibles en annexe.

SMR

- **Exclure les résumés hebdomadaires de sortie anonymes (RHA)** qui ne seront pas valorisés, (i.e. avec une erreur de groupage) via la table T_SSRaaB
  - Variable disponible depuis 2013 : *WHERE GRG_GME NOT LIKE '90%'* ;
- **Exclure les séjours ou parties de séjours** qui ne seront pas valorisés (i.e. avec une erreur de groupage) via la table T_SSRaaGME
  - Variable disponible depuis 2012 : *WHERE GME_COD NOT LIKE '90%'* ;
  - et après 2012 : *WHERE GME NOT LIKE '90%'* ;
- **Exclure les « faux » RHA générés automatiquement** pour les besoins de facturation (déjà exclus via les filtres précédents) via la table T_SSRaaB
  - Variable disponible depuis 2015 : *WHERE TYP_GEN_RHA IN ('0', '4')* ;
- **Exclure les RHA de l'année précédente** (i.e. RHA répétés dans le PMSI de l'année N pour les séjours non clos et non valorisés en N-1) via la table T_SSRaaB : *WHERE RIGHT(MOI_ANN, 4) = annee* ;
- **Exclure les séjours associés à des clés de chaînage incorrectes sur les informations des bénéficiaires** via les variables codes retours de la table T_SSRaaC :
  - Variables disponibles depuis 2005 : *WHERE NIR_RET = '0' AND NAI_RET = '0' AND SEX_RET = '0' AND SEJ_RET = '0' AND FHO_RET = '0' AND PMS_RET = '0' AND DAT_RET = '0'* ;
  - Variables disponibles depuis 2013 : *AND COH_NAI_RET = '0' AND COH_SEX_RET = '0'* ;
- Dans ce champ d'activité, **les ES APHP, APHM et HCL n'ont pas remonté leur activité en double sur leur FINESS géographique.**

Appliquer des filtres dans son extraction (données du PMSI) (5/5)

Séjours : il est recommandé **d'appliquer systématiquement plusieurs filtres**

RIM-P

- **Exclure les séquences indiquées comme « sortie d'essai » jusqu'en 2016** car elles ne sont pas considérées comme des hospitalisations via la table T_RIPaaRSA : *WHERE SEQ_IND <> 'E'* **;**
- **Exclure les « faux » résumés de sortie anonymes (RSA) générés automatiquement** pour les besoins de facturation via la table T_RIPaaRSA
  - Variable disponible depuis 2015 : *WHERE TYP_GEN_RSA = '0'* ;
- **Exclure les séjours associés à des clés de chaînage incorrectes sur les informations des bénéficiaires** via les variables codes retours de la table T_RIPaaC :
  - Variables disponibles depuis 2007 : *WHERE NIR_RET = '0' AND NAI_RET = '0' AND SEX_RET = '0' AND SEJ_RET = '0' AND FHO_RET = '0' AND PMS_RET = '0' AND DAT_RET = '0'* ;
  - Variables disponibles depuis 2013 : *AND COH_NAI_RET = '0' AND COH_SEX_RET = '0'* ;
- Dans ce champ d'activité, **les ES APHP, APHM et HCL n'ont pas remonté leur activité en double sur leur FINESS géographique.**

Des programmes SAS illustrant ces éléments sont disponibles en annexe.



4. Exercice pratique

Exercices pratiques - Introduction



- **Deux exercices pratiques** sont proposés :
  - Pour être réalisés de façon autonome ;
  - Afin de permettre aux utilisateurs de s'exercer à manipuler une base de données synthétiques de la base principale du SNDS.

- **Pour réaliser ces exercices**, un jeu de données synthétiques de la base principale du SNDS est mis à disposition sur Gitlab, restitué au format dans le cadre d'un accès permanent.

- **La correction de chaque exercice** est fournie pour permettre aux utilisateurs de s'auto-évaluer et de progresser en autonomie.



Vous pouvez poser vos questions sur le forum de la communauté d'entraide SNDS

Exercices pratiques - Le jeu de données synthétiques (1/2)

Le jeu de données synthétiques a été généré **à partir d'un programme SAS®** conçu pour **reproduire la structure et la logique des données issues de la base principale du SNDS**. Ce programme, développé par Jérôme Brocca à partir de son expertise métier, n'utilise aucune extraction des données source à partir de la base principale du SNDS. Les données synthétiques obtenues reprennent **la même organisation (librairies, tables, variables, codage ...) et ont la même cohérence que les données source** disponibles sur la plateforme SNDS dans le cadre d'un accès permanent.

Ce jeu de données contient :

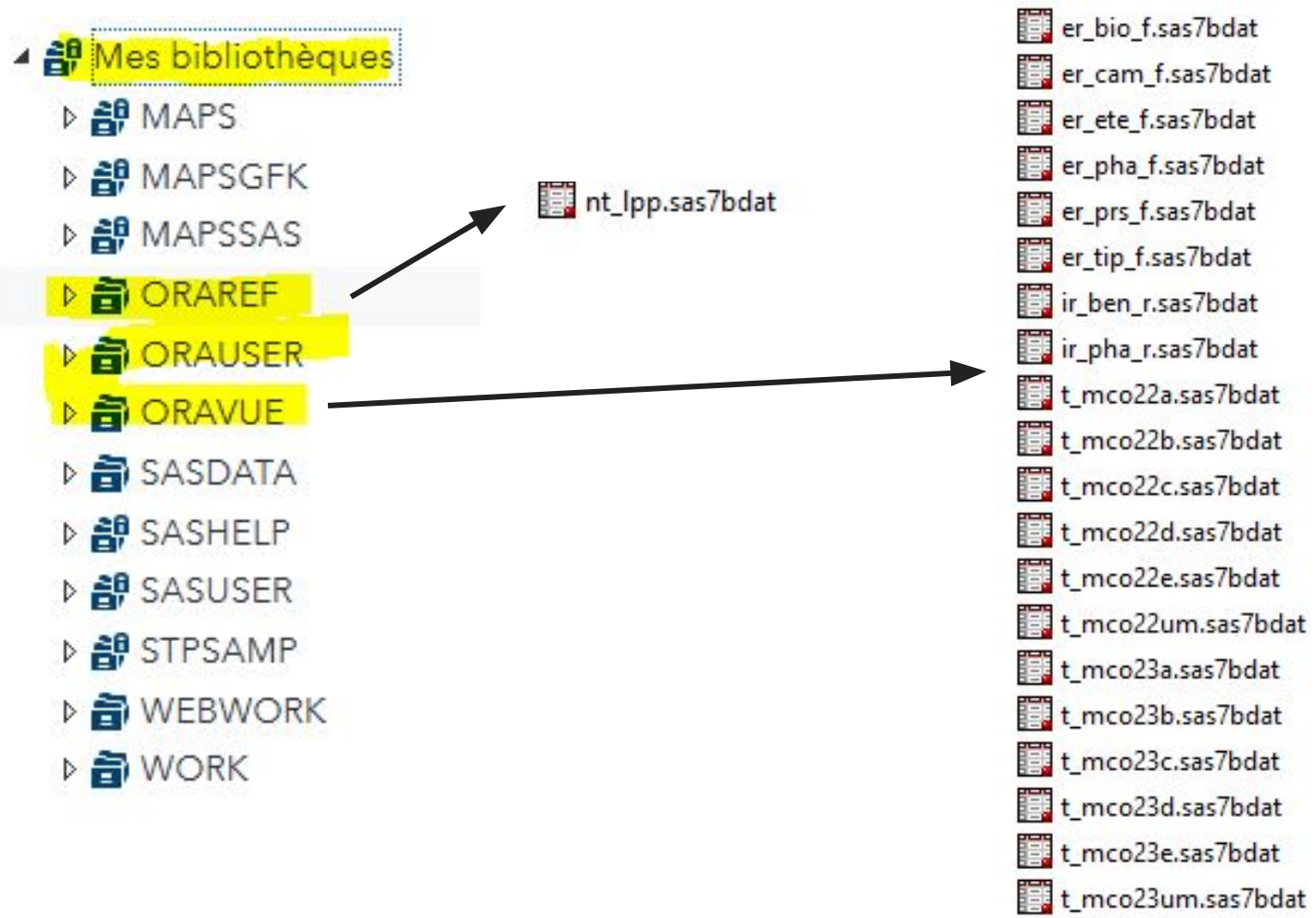
1 092 sujets distincts

Des données de remboursement de soins de ville DCIR (médicaments, consultations, procédures ...) sur la période 2023-2025

Des données de séjours hospitaliers du PMSI MCO sur la période 2022-2023

Exercices pratiques - Le jeu de données synthétiques (2/2)

Pour les deux exercices, les librairies et tables *.sas7bdat* suivantes ont été générées :



Jeu de données et exercices pratiques (1/2)

Les attendus

Exercice pratique n°1 :

Calculer par type d'exécutant (PS/Etablissement), spécialité exécutant et prestation, **le nombre d'actes de téléconsultation retrouvés dans le DCIR en 2023** :

- ➔ Avec l'ensemble des remboursements
- ➔ En ne prenant que les soins de ville (élimination des établissements publics)

Indications



Prestations sélectionnées

1056	JC	TELECONSULTATION GENERALISTE IVG
1057	JCS	TELECONSULTATION SPECIALISTE IVG
1086	TFS	TELESOINS SF
1096	TTE	TELECONSULTATION MEDECIN TRAITANT AVEC EHPAD
1191	TC	TELECONSULTATION TOUTES SPECIALITES
1192	TCG	TELECONSULTATION GENERALISTE
1046	TCS	TELECONSULTATION SPECIALISTE

Les résultats

➔ Nombre de téléconsultations en 2023

Type exécutant	Spécialité exécutant	Prestation	Nombre de téléconsultations
Etablissement	1	1192	99
Etablissement	33	1191	44
PS Libéral	1	1192	108
PS Libéral	12	1191	61
PS Libéral	21	1086	86
TOTAL	-	-	398

Jeu de données et exercices pratiques (1/2)

Les attendus

2

Exercice pratique n°2 :

Rechercher le nombre de bénéficiaires ayant eu un séjour avec diabète en 2023 :

➔ Au moins un séjour PMSI MCO avec présence d'un diagnostic de diabète (DP ou DR du séjour, DP ou DR de RUM, DA)

Préciser le nombre de patients avec des clés de chaînage OK

Indications



Diagnostics du diabète : Codes CIM10 : E10 à E14

Les résultats

➔ Répartition des patients pour séjour MCO diabète

Nombre total bénéficiaires topés	Dont nombre bénéficiaire chaînage OK	Nombre bénéficiaires hospitalisés motif principal	Dont nombre bénéficiaires hospitalisés motif principal chaînage OK
456	454	276	275

PLATEFORME DES DONNÉES DE SANTÉ
FRENCH HEALTH DATA HUB

Suivez-nous sur les réseaux sociaux !