Thème : Appariement

Fiche : APPARIER DES DONNEES EXTERNES AUX TABLES DU SNDS

Objectif

Comprendre le principe d'appariement de données externes avec les données du SNDS et les exploiter correctement

Public ciblé

Toute personne ayant accès aux données du SNDS, produits SAS

1- Les appariements dans le SNDS - Définition

Le principe d'un appariement consiste à associer des données externes au SNDS avec des données internes via des informations communes (informations sur les bénéficiaires, informations sur les Professionnels de Santé…). Il existe deux types d'appariements :

- Les appariements directs (déterministes)
- Les appariements indirects (probabilistes)

Ces appariements nécessitent une autorisation de la CNIL au préalable.

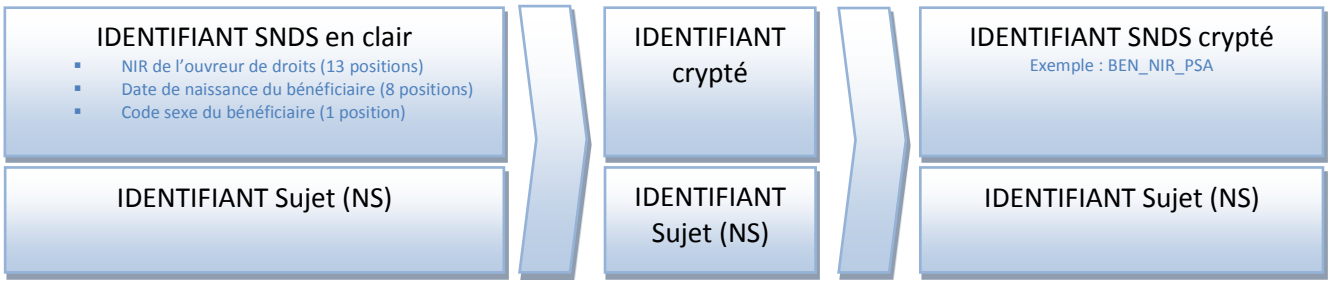
Exemple 1 : Un chercheur souhaite obtenir des informations sur une typologie de patients pour lesquels il a leur NIR. Il va entreprendre les démarches auprès de l'équipe DEMEX de la CNAM et va réaliser une demande de pseudonymisation de leurs données (en utilisant la procédure nécessaire) à partir des numéros de sécurité sociale, des codes sexe et des dates de naissance fournis. Il aura en sa possession une correspondance entre ses données et un identifiant sujet (NS) des patients de son étude. L'équipe DEMEX de la CNAM récupérera l'identifiant crypté, présent dans le SNDS (à aucun moment, ils ne pourront remonter au numéro de sécurité sociale en clair) sur lequel ils pourront extraire les données de consommation et les livrer, associées à l'identifiant sujet (NS). C'est un **appariement direct**.

Exemple 2 : Un chercheur souhaite obtenir des informations sur 1000 patients d'une liste de 10 médecins sélectionnés (100 patients par médecin). Il fournit une liste de variables communes pour tous les patients de son étude (par exemple : âge, sexe, prescription médicaments, dates de soin…). Chaque patient est reconnu au niveau du chercheur par un identifiant sujet (NS).
L'équipe DEMEX va donc récupérer les informations nécessaires et rechercher dans les tables du SNDS les patients pouvant correspondre de manière probabiliste aux critères.
Une extraction des données SNDS sera alors réalisée pour ces patients. Ils seront reconnus par le chercheur sur leur identifiant sujet (NS). C'est un **appariement indirect**.

2- Les appariements directs

2.1 Appariements sur données identifiantes

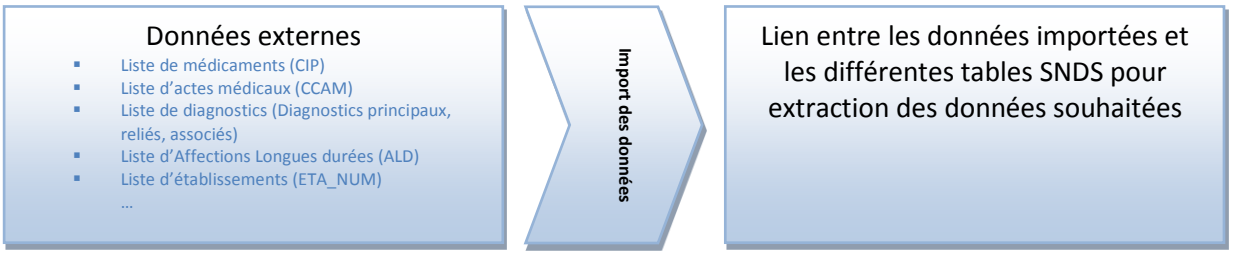
L'appariement direct consiste donc à faire le lien entre des patients connus du demandeur (par leur NIR/date de naissance et sexe) et leurs données de consommation de soins du SNDS. Afin de faire ce lien, il faut donc impérativement que l'identifiant SNDS soit donné.
La procédure passe par un cryptage irréversible des informations Numéro de sécurité sociale-Date de Naissance-Sexe pour transformer l'identifiant SNDS en clair en identifiant SNDS crypté (BEN_NIR_PSA). Cette procédure se fait via l'outil SAFE et se décompose de la manière suivante :



Le commanditaire fournit, en plus des informations de l'identifiant SNDS en clair, un numéro sujet (par exemple NS_000001 -> NS_999999) qui va persister au cryptage pour être associé à l'identifiant SNDS crypté final. Les données du SNDS seront extraites sur ce dernier et livrées sur l'identifiant sujet.
Le commanditaire pourra faire son étude sur sa population et répondre à sa problématique.

2.2 Appariements sur données anonymes

Un autre appariement direct pourrait consister à importer des données dans le portail SNDS (en utilisant l'outil d'import/export) comme par exemple une liste de médicaments, d'actes médicaux, des diagnostics, de départements... Un lien pourrait être réalisé entre ces données et celles du SNDS pour ensuite extraire les données des bénéficiaires correspondants. Cet appariement est équivalent à une sélection sur critères.

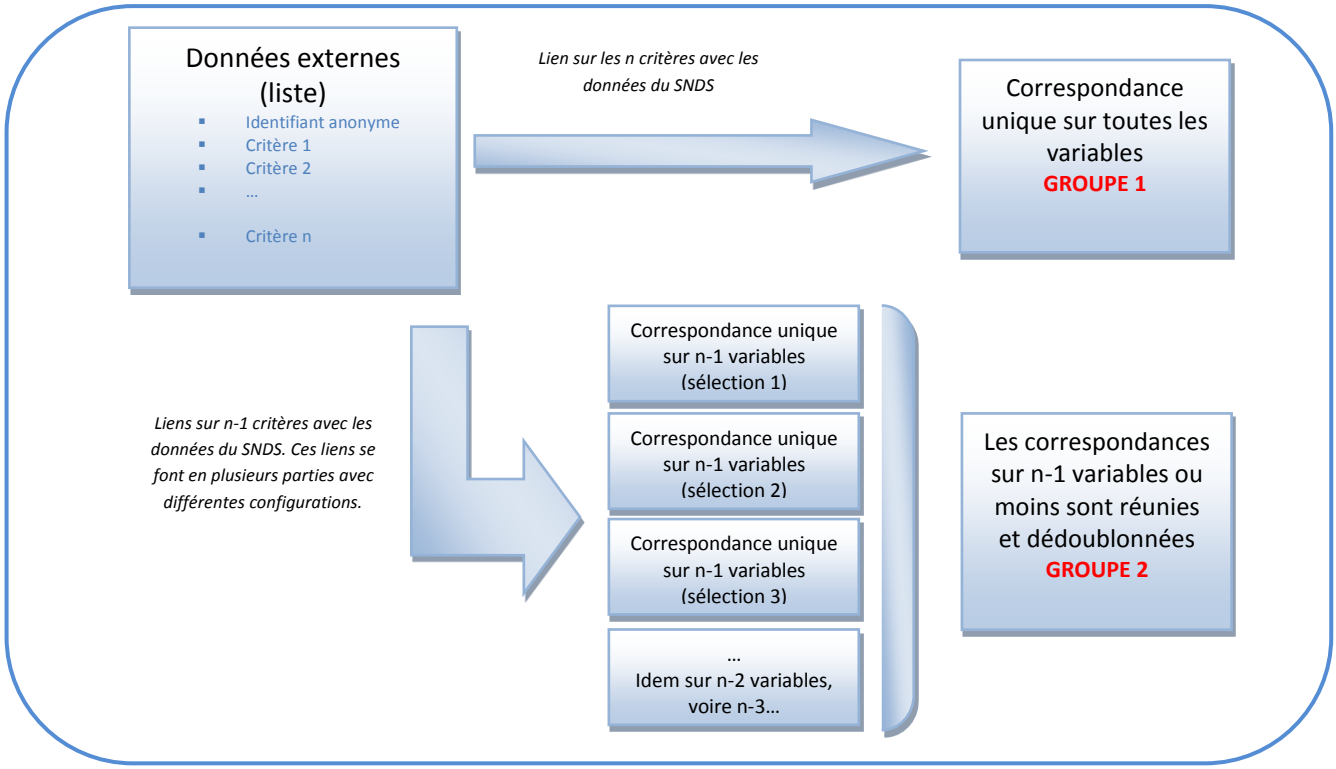


NB : Attention, l'import/export de données ne doit pas concerner des données identifiantes.

3- Les appariements probabilistes

L'appariement probabiliste consiste à sélectionner des patients en fonction d'informations définies sans avoir d'identifiant en clair. Ces patients devront correspondre de manière probabiliste aux informations souhaitées pour que la correspondance soit la plus optimale possible.

La procédure de correspondance se fait généralement en plusieurs étapes, par tâtonnement, en jouant sur les différentes variables.
Dans un premier temps, une liste de patients est envoyée avec des informations concrètes. Aucune information directement identifiante du bénéficiaire n'est contenue dans cette liste.
Par exemple, la liste peut contenir un identifiant sujet et n critères de sélection (sexe, département de résidence, date de soin, code diagnostic principal, régime d'affiliation, année de naissance…).
Dans un second temps, la liste est jointe aux tables du SNDS sur 100% des critères pour essayer d'obtenir une unicité de patients correspondant parfaitement à **TOUS** les critères (1 identifiant sujet = 1 identifiant crypté). C'est le **GROUPE 1** du schéma ci-dessous. Il peut correspondre à 25%-50%-75%-100% de la liste initiale.
Ces patients ont donc parfaitement été retrouvés. Pour ceux qui n'ont pas de correspondance, il faut renouveler l'opération en enlevant un critère (puis ensuite un deuxième si nécessaire…) moins discriminant. Cette étape peut se faire en plusieurs parties et va permettre de sélectionner de manière probabiliste de nouveaux patients dont la correspondance sera également unique. C'est le **GROUPE 2** du schéma.
Les 2 groupes réunis forment la sélection finale.
Le principe peut évidemment être réalisé avec davantage de groupes, selon les échanges avec le commanditaire de l'étude.
Un appariement sera alors considéré comme réussi si au moins 90% des patients initiaux seront retrouvés.



Un travail d'ajustement des formats de variables peut être à prévoir car les variables des fichiers sources ne sont pas toujours au même format que les données du SNDS.
Toutefois, la sélection est rendue plus facile lorsque le nombre de variables transmises est important (avec des variables le plus discriminantes possibles sur le patient). Cependant, une correspondance multiple peut être détectée et sera exclue des groupes. Le 100% est donc difficilement atteignable.
La qualité des données initiale est donc primordiale pour un appariement probabiliste réussi.