Standardisation de l'utilisation des données du SNDS à des fins de recherche médicale : présentation d'un modèle de données optimisé centré sur le patient

N.Thurin[1], R. Lassalle[1], J. Jové[1], R. Sylvestre[1], J. Dupont[1], C. Lacueille[1], P. Blin[1], N. Moore[1], C. Droz-Perroteau[1]

[1] Bordeaux PharmacoEpi, INSERM CIC1401, Université de Bordeaux, Bordeaux, France

Introduction

- **Le SNDS est un outil puissant et incontournable en pharmaco-épidémiologie permettant**
  - Une analyse d'impact (bénéfices/risques) des médicaments en conditions réelles,
  - L'évaluation de parcours de soin des patients,
  - Le suivi de l'évolution des maladies et des traitements.
- **Le SNDS a une architecture centrée sur la prestation (DCIR) et le séjour (PMSI, Figures 1 et 2)**

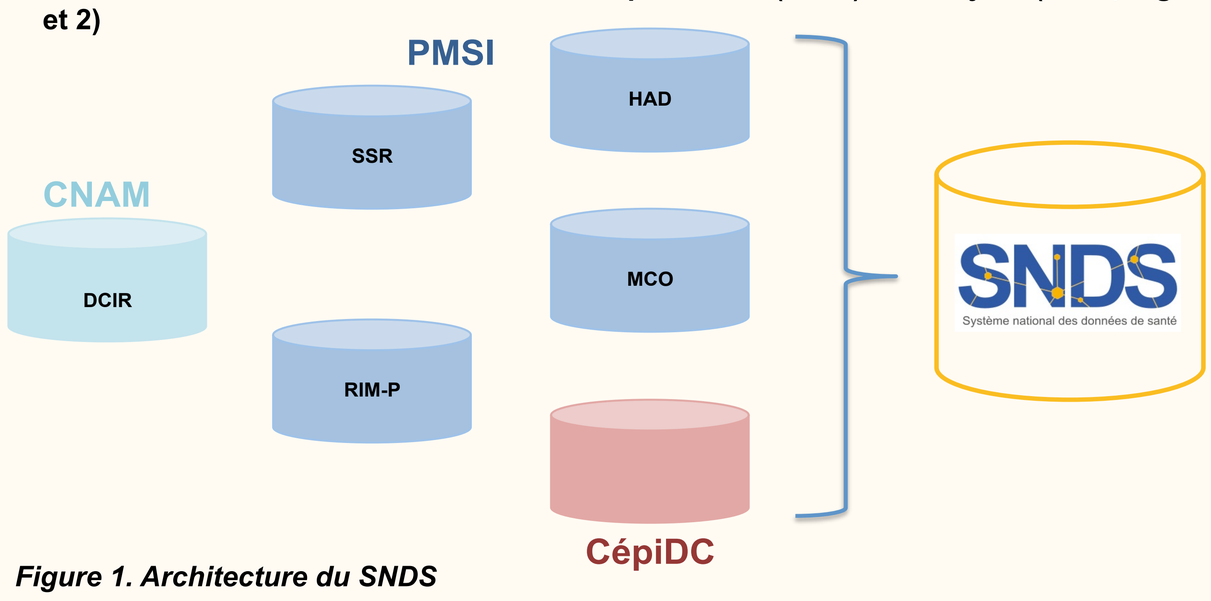
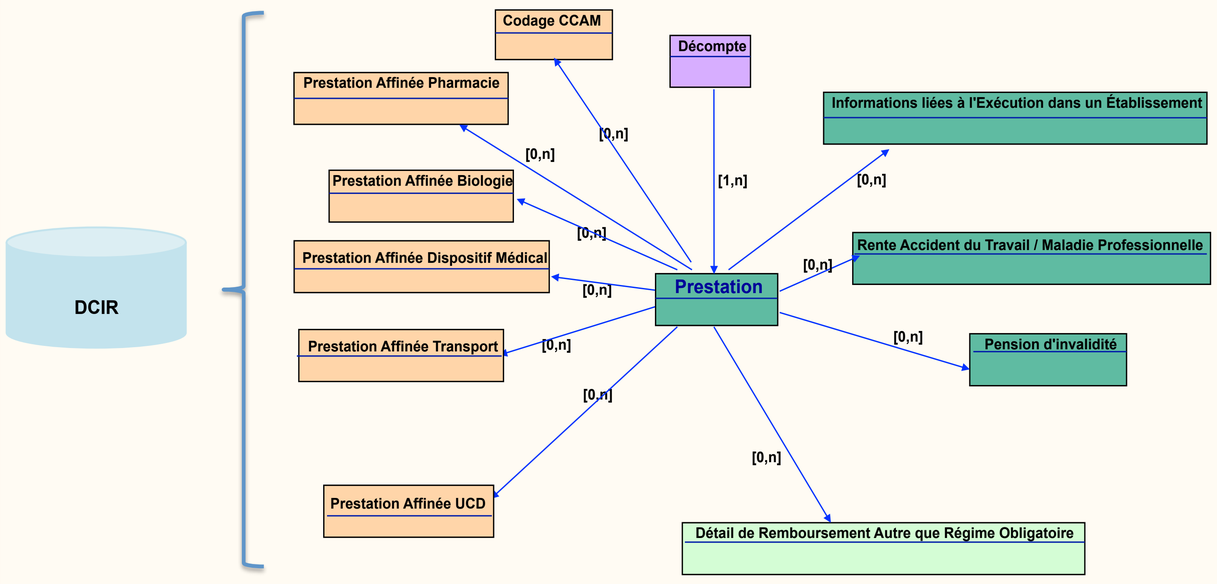


Figure 1. Architecture du SNDS



Source : http://open-data-assurance-maladie.ameli.fr/wiki-sniiram/

Figure 2. Architecture du DCIR centrée sur la prestation

- **L'architecture du SNDS a pour conséquence de rendre complexe son usage à des fins de recherche (Tableau 1)**
  - L'unité statistique est la prestation ou le séjour plutôt que le patient,
  - Le nombre de tables annuelles et de variables est très important et croissant,
  - Cela entraine une certaine « lourdeur » dans les traitements : nombre élevé de clés de jointure, grand volume de données et tables à manipuler, temps machine important.

Univers	Nombre de tables/an[1]		Nombre de variables/an[1]	
	Min	Max	Min	Max
DCIR	12	13	375	394
MCO	19	54	342	859
SSR	14	27	260	519
RIM-P	6	16	92	361
HAD	6	22	114	437
CépiDC[2]	-	2	-	61
Total	**57**	**134**	**1183**	**2631**

[1] 2006 – 2016 [2] 2013 – 2015

Tableau 1. Volume annuel de tables et variables au sein du SNDS

Objectifs

Construire, à partir des données du SNDS, un modèle simplifié et optimisé pour conduire des études de pharmaco-épidémiologie centrées autour du patient, avec un nombre de tables et de variables restreint.

Méthodes

- **Définition du modèle de données relationnel par une équipe pluridisciplinaire**
- **Rédaction du cahier des charges (ETL « Extract Transform and Load »)**
  - Sélection des variables pertinentes (exposition médicamenteuse, pathologies, indicateurs de soins…),
  - Description des tables,
  - Cartographie des variables sources vers le modèle final,
  - Création du thésaurus associé.
- **Programmation et exécution sur données « test » (SAS®)**
- **Validation des programmes et des tables créées**

Résultats

- **Modèle de données final centré autour du patient (Figure 3)**

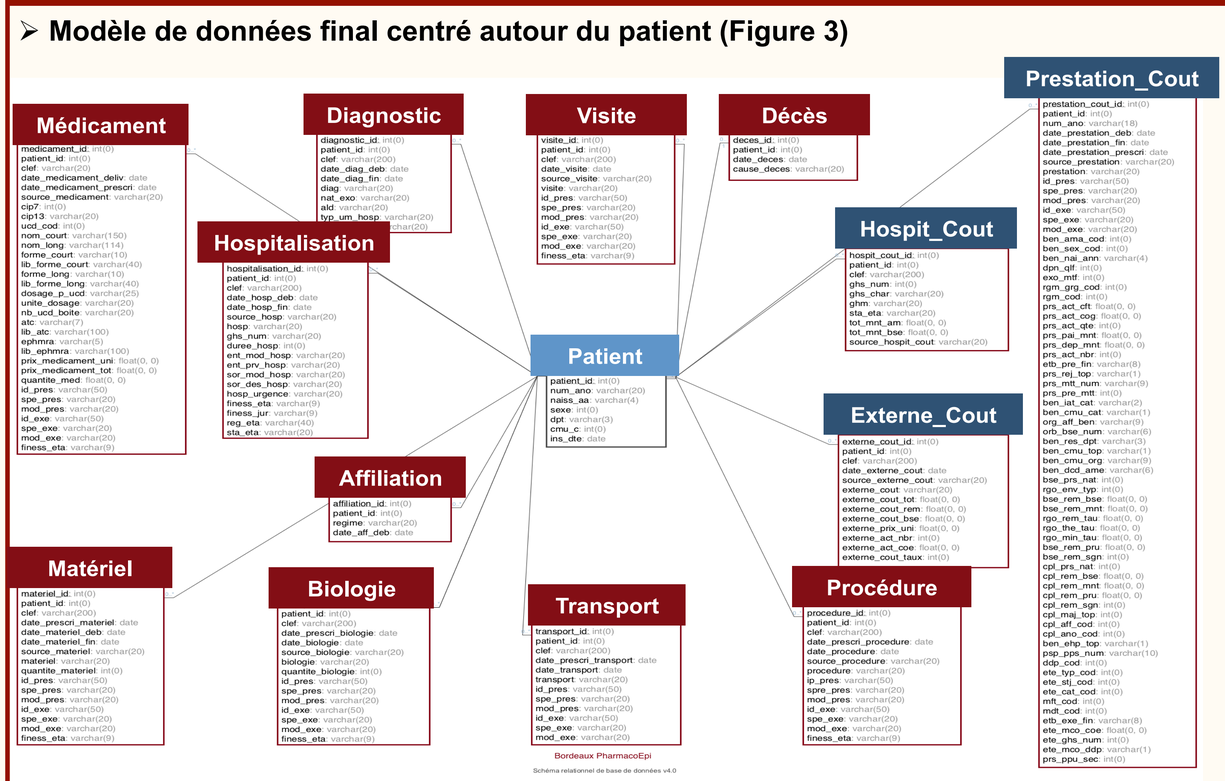


Figure 3. Architecture du SNDS centrée sur le patient

- **Structure des données, exemple de la table « Diagnostic » (Tableau 2)**
  - Données, provenant de 13 tables/année dans 5 des 6 univers (hors CépiDC), synthétisées en une seule table,
  - Clé patient unique, code CIM-10 (diagnostic, ALD) réduit à 1 seule variable

Nom Variable	Libellé Variable	Type Variable	Format	Précision
ID	Identifiant du diagnostic (clé primaire)	number	11.	
PATIENT_ID	Identifiant du patient (clé étrangère)	number	11.	
CLEF	Clef (pour relier avec table HOSPITALISATION, HOSPIT_COUT et autres données hospitalières)	char	$200.	
DATE_DIAG_DEB	Date de début (séjour hospitalier / mise en ALD ou invalidité ou AT ou MP)	date	8.	
DATE_DIAG_FIN	Date de fin (séjour hospitalier / mise en ALD ou invalidité ou AT ou MP)	date	8.	
DIAG	Code CIM10 Diagnostic hospitalier / ALD / invalidité / AT / MP	char	$20.	Vocabulaire CIM10
NAT_EXO	Nature d'exonération	char	$20.	Vocabulaire NAT_EXO
ALD	Code ALD	char	$20.	Vocabulaire TYP_ALD
TYP_UM_HOSP	Type d'autorisation d'unité médicale	char	$20.	Vocabulaire TYP_UM
SOURCE_DIAG	Source du diagnostic	char	$20.	Vocabulaire SOURCE_DIAG

Tableau 2. Exemple de la table « diagnostic »

- **Pour chaque table**
  - Indentification de la source de provenance de chaque donnée,
  - Calcul de variables dérivées (passage aux urgences, soins intensifs, CMU-c…),
  - Possibilité de relier des données d'un même univers présentes dans différentes tables,
  - Ajout dans la table des coûts hospitaliers « Hospit_cout » du montant environné du séjour provenant des Etudes Nationales de Coûts (ENC) annuelles de l'ATIH.
- **Exemple de l'étude CAMERRA : cohorte 2012 – N = 913 358 patients, extraction de 10 années de soins (Tableau 3) :**

	Modèle initial	Modèle simplifié	Modèle simplifié sans la table "Prestation_cout"
Nombre de tables	973	**14**	**13**
Volume des données SAS®	473 Go	**389 Go (82%)**	**52 Go (11%)**

Tableau 3. Etude CAMERRA, application du modèle simplifié

→
- **Gain de tables et de volume de données SAS®**
- **Table « Prestation_cout » encore volumineuse**

Conclusion

- **Cette standardisation de l'utilisation des données du SNDS a permis**
  - **D'optimiser le travail des statisticiens**
    - **Standardisation des tables,**
    - **Diminution de la variabilité inter-programmeurs,**
    - **Gain notable en terme de temps,**
    - **Meilleure reproductibilité des résultats et qualité scientifique.**
  - **D'optimiser l'exploitation des données**
    - **Utilisation simplifiée d'autres outils (macro SAS® HdPS, packages R…),**
    - **Simplification de la présentation et explication de la structure des données,**
    - **Programmes de recherche internationaux multi-pays, multi-bases de données : interconnexion optimisée avec les plateformes existantes (exemples : Aetion, Pharmo, OMOP-OHDSI…).**

2ème congrès "Le SNDS pour la recherche en santé"
24 et 25 juin 2019, Rennes – France