



HAL
open science

SNDS: contexte, contenu et usages de la base de l'assurance maladie

Thomas Guyet

► **To cite this version:**

Thomas Guyet. SNDS: contexte, contenu et usages de la base de l'assurance maladie. Symposium MADICS, GdR MADICS, May 2024, Blois, France. hal-04593382

HAL Id: hal-04593382

<https://inria.hal.science/hal-04593382>

Submitted on 29 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

SNDS: contexte, contenu et usages de la base de l'assurance maladie

T. Guyet – Inria/AlstroSight
thomas.guyet@inria.fr



Atelier TIDS – Symposium MADICS
29/05/2024

SNIIRAM/SNDS: une base de données très convoitée

- SNDS = Bases de données de santé nationales
- **Intérêt fort à exploiter cette base de données** pour la recherche médicale et les pouvoirs publics
 - accélérer le développement de la recherche médicale
 - faciliter les études et le suivi (pharmaco-)épidémiologique (e.g. Mediator) [MBL+24]
 - suivre et prévoir des dépenses de santé
 - gérer les ressources de soins du territoire
- Objectivement **intéressante au niveau international** : grande couverture d'une population large
- Soulève des **défis informatiques importants**
 - et donc aussi des intérêts pour la communauté informatique
 - opportunité des "accès permanents"
 - ↔ contraint par des restrictions d'accès (dont RGPD)

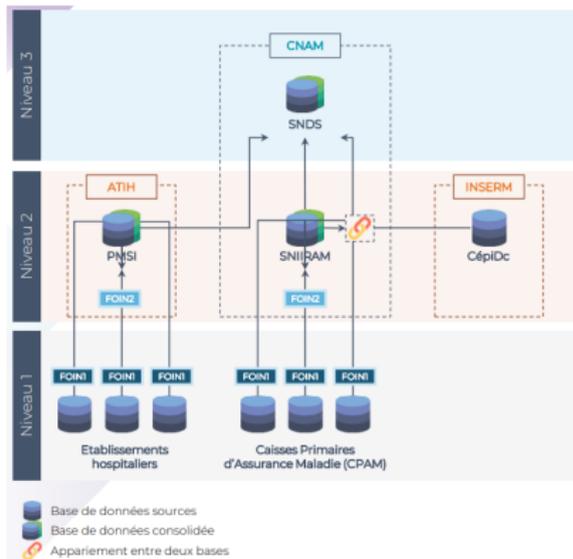
⇒ Est-ce que cela je peux accéder aux données du SNDS ?

⇒ Quelles données vais-je y trouver ?

SNDS: Quelques repères historiques (source IRDES)

- 1977 : création du **SNIR** (Système National Inter-Régimes) par l'Assurance maladie.
- 1992 : début du projet de Médicalisation des Systèmes d'Information (**PMSI**)
- 1999 : création du Sniiram (Système national d'information inter-régimes de l'Assurance maladie) par la loi de financement de la sécurité sociale pour 1999
- 2010 : chaînage effectif PSMI/SNIIRAM
- 2015 : début projet ANSM/PEPS
- 2016 : loi de modernisation du système de santé instaure le Système National des Données de Santé (SNDS)
- 2018/03 : Rapport Villani sur le développement de l'intelligence artificielle
- 2018/10 : Mission de préfiguration du Health Data Hub
- 2019 : La Loi du 24 juillet 2019 relative à l'organisation et la transformation du système de santé crée Health Data Hub, plateforme nationale des données de santé

Données de l'assurance maladie : SNDS/SNIIRAM I



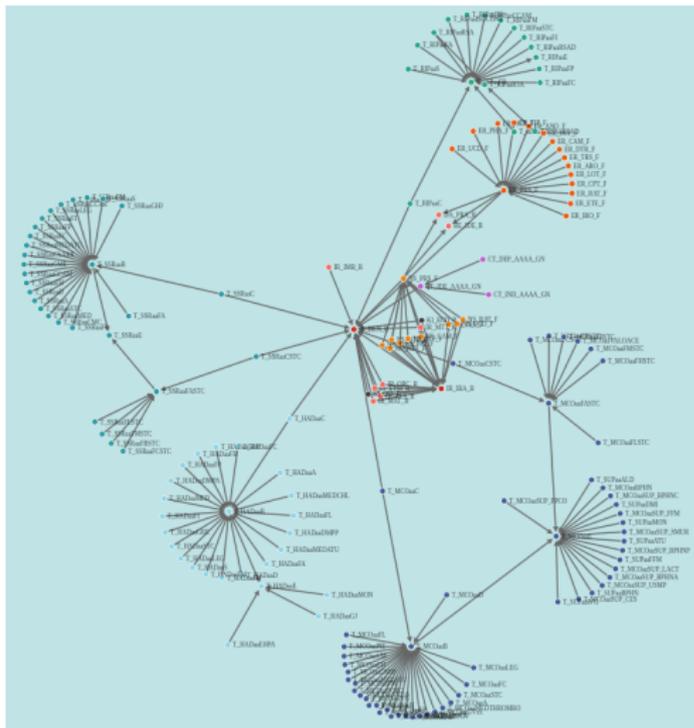
- SNIIRAM: Les Caisses Primaires d'Assurance Maladie remontent l'ensemble des informations issues des remboursements à la CNAM.
- PMSI: Chaque établissement de santé enregistre chacun des séjours hospitaliers
 - ATIH: responsable de Coder, Recueillir, Analyser, Restituer et Diffuser l'information hospitalière
- CépiDc: Le CépiDc de l'Inserm gère la Base de Causes Médicales de Décès (BCMD).
- EGB: Échantillon généraliste des bénéficiaires

Données de l'assurance maladie : SNDS/SNIIRAM II

Principales informations du SNDS

- **Bénéficiaire** (sexe, mois, année et rang de naissance, lieu de résidence, régime, CMU, aide à la complémentaire santé) ;
- **Professionnels de santé** (spécialité, mode d'exercice, sexe, âge, département d'implantation)
- Pathologies, notamment les **ALD** et les diagnostics des séjours hospitaliers (PMSI) (**motif d'hospitalisation**) ;
- Dépenses et remboursements (prestations en soins de ville, en établissements de santé, et montants associés)
- Consommations de **soins de ville** (consultations, actes techniques...)
- **Délivrances de produits de santé** (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** (code CIM)

Données de l'assurance maladie : SNDS/SNIIRAM III



La base de données ...

- 692 tables
- plusieurs milliers d'attributs
- ~ 500 tables de nomenclature, nomenclature = codage "standard"
 - CIP/ATC: code médicaments
 - CIM: code diagnostic
 - NABM: actes de biologies
 - ...
- organisation en étoile(s)

<https://health-data-hub.shinyapps.io/dico-snds/>

DCIR

Les référentiels de DCIR s'écrivent de la forme : **IR_XXX_R**

IR_BEN_R

Référentiel Bénéficiaires

IR_UMB_R

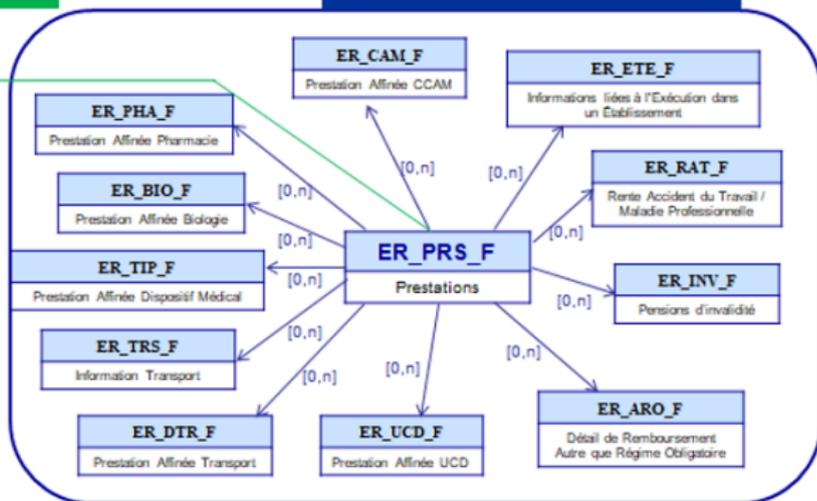
Référentiel Médicalisé des Bénéficiaires

informatif sur les bénéficiaires

Lien via {BEN_NIR_PSA ; BEN_RNG_GEM}

Les 12 tables DCIR s'écrivent de la forme :
ER_XXX_F (12 tables annuelles)

Flux de consommation du bénéficiaire



Quelques caractéristiques des données

Quelques caractéristiques [ZCG21]

- Information temporelle riche (granularité journalière)
 - Parcours de soins
 - Activité des praticiens
- Information géographique sur les praticiens et les structures
- Données structurées
 - Beaucoup de codifications
- Couverture des nombreuses activités de santé

Explorer plus encore le SNDS ...

- Beaucoup d'information disponibles sur la documentation technique
 - <https://documentation-snds.health-data-hub.fr/>
 - <http://docs.collectif-cocoa.org/>
 - Documentation CNAM (accès restreint)
 - articles scientifiques [TRC⁺17, MBL⁺24]
 - Base de données synthétiques techniquement représentative [Guy21]
 - <https://zenodo.org/record/4659983>
 - Générateur:
https://gitlab.inria.fr/tguyet/medtrajectory_datagen
- disponible et flexible

Exemple de base de données synthétiques

Retour à la réalité ...

Ce que ne contient pas la base

- **Informations masquées** car pas données observées
 - Les soins non-remboursés (e.g. para-pharmacie)
 - Les prestations à l'hôpital
- Les posologies
- Uniquement les actes mais **pas les résultats** d'examens
 - pas d'images
 - pas de comptes rendus hospitaliers
- **Pas de diagnostics** (hors hôpital)
- L'observance ... et les **prescriptions non-délivrées**

Comparatif avec les données hospitalières

EDS

Données disponibles variées et riches

- Données structurées et non-structurées
- Informations “médicales” riches
- ... mais données non-structurées pas toujours facilement accessibles !

Couverture de la base

- Population limitée
- Visibilité uniquement de l'activité intra-hospitalière
- Temporalité fine

SNDS

Couverture large, exhaustivité

- Des prestations de soins
- De la population (avec une population importante)
- Temporalité longue
- Fiabilité des données (administratives)

Informations médicale limitée

- Absence de diagnostics
- Complexité technique

⇒ la complémentarité motive l'intérêt de l'appariement de ces bases [JCB⁺22] ... *mais c'est la situation la plus difficile à mettre en*

Retour à la réalité ... (2)

Des données complexes

- **Variables parfois complexes** (héritage d'empilement)
 - Clés de jointure à 9 variables → attention aux temps de calcul!
 - Beaucoup d'exceptions
 - **Longitudinalité pas si simple**
 - Modification dans le temps des codages (et des champs remboursés)
 - Pas de code unique toute sa vie (enfant->adulte)
 - **Homogénéisation des données** d'une population
 - Des régimes de remboursements différents
 - ...
- Nécessite une bonne connaissance du SNIIRAM et du système de santé français

Simplifier l'utilisation de la base de données

- Utilisation d'un schéma RDF [RDLM18, BGD⁺21] (+ en cours)
- Le HdH propose une conversion vers un schéma OMOP

Écart sémantique

Changement des usages des données

- Usage primaire : informations collectées pour des objectifs (administratifs)
 - Usage secondaire : utilisation souhaitée pour un autre usage (médical)
- spécificité par rapport à la collecte de données : changement épistémologique important ... fortement lié aux méthodes d'analyse de données / IA

Écart sémantique

- Utilisation d'une information dans un champs sémantique différent de celui pour laquelle elle a été collectée
 - Exemple : identification d'une *pathologie* à partir des données du SNDS (*délivrances de soins*)
- ⇒ nécessite de raisonner à partir des données

Accès aux données (vue simplifiée)

Quatre grands types d'accès aux données

- 1 Chargement local de données : nécessite un très rare status de "SNDS-fils"
- 2 Accès par le portail de la CNAM à la base
 - SAS Guide ou Rstudio
 - nécessite des formations avant accès
- 3 Accès à l'**EGB** (Échantillon Généraliste des Bénéficiaires)
- 4 Accès par la plateforme du HdH (*voir Tom*)

Données accédées par statut

- Accès permanents (dont organismes de recherche): les données accédées changent en fonction des organismes (accès CNAM ... plateforme HdH ?)
 - Inria requiert des validations par le comité éthique des demandes d'accès (sur projet)
- Accès non-permanents : nécessite un projet évalué au CESREES

Appariement des données

Appariement de données [JCB⁺22]

- Mise en commun de deux sources de données
 - SNDS + données spécifiques
 - données spécifiques sur des patients : imagerie, données hospitalière, etc.
- ⇒ objectif d'enrichir des analyses de son jeu de données par l'enrichissement des données avec celles du SNDS
- Deux types d'appariement : par tier de confiance vs appariement statistique

Limites techniques et administratives des appariements

- Nécessite une demande autorisation spécifique (même avec un accès permanent)
- Nécessite une infrastructure pour recueillir les données à appairer

Recherche avec le SNDS

Recherche médico-sociale

- Épidémiologie : analyse retrospective de cohorte
 - Analyse de l'effet des produits de santé
 - Analyse des trajectoires de soins
 - Économie de la santé : analyse de la consommation des soins
- ⇒ les infrastructures d'accès de la CNAM sont faites pour ce type de recherche

Recherche en informatique : de nombreux défis

- De nombreux défis pour améliorer l'exploitation de la base pour les besoins ci-dessus
 - Par exemple:
 - Enrichissements sémantiques : représentation, requêtage, etc.
 - Analyse de données massives : stockage, visualisation
 - Machine learning : analyse de trajectoires, prédiction des soins, etc.
- ⇒ Les modalités d'accès 2 et 3 ne permettent pas de mener facilement ces travaux

Conclusion

- SNDS: très convoitée pour la recherche
 - en santé
 - en informatique (et plus spécifiquement ML)
- Que vais je trouver dans le SDNS ?
 - ↪ données médico-administratives de remboursement de soins
 - ↪ contenu technique et ne répondant qu'à certains besoins
 - ⇒ possibilité d'explorer une base de données synthétiques "réaliste"
- Est-ce que je peux accéder aux données du SNDS ?
 - ↪ oui ... avec différentes difficultés administratives et techniques

References |



Johanne Bakalara, Thomas Guyet, Olivier Dameron, André Happe, and Emmanuel Oger, *An extension of chronicles temporal model with taxonomies -Application to epidemiological studies*, HEALTHINF 2021 - 14th International Conference on Health Informatics (online, France), February 2021, pp. 1–10.



Thomas Guyet, *Génération d'un snds synthétique à partir de données ouvertes*, Actes de la conférence Extraction et Gestion des connaissances (Démon), 2021, pp. 1–2.



C Jean, P Caillet, S Bréant, C Daniel, C Hassen-khodja, E Paillaud, E Audureau, and F Canouff-poitrine, *Trajectoires de soins des patients âgés atteints de cancer: chaînage de la cohorte clinique elcapa avec l'entrepôt de données de santé de l'ap-hp (projet elcapa-eds)*, Revue d'Épidémiologie et de Santé Publique 70 (2022), S97–S98.



Olivier Maillard, René Bun, Moussa Laanani, Amandine Verga-Gérard, Taylor Leroy, Nathalie Gault, Candice Estellat, Pernelle Noize, Florentia Kaguelidou, Agnès Sommet, Maryse Lapeyre-Mestre, Annie Fourier-Réglat, Alain Weill, Catherine Quantin, and Florence Tubach, *Use of the french national health data system (snds) in pharmacoepidemiology: A systematic review in its maturation phase*, Therapies (2024).



Yann Rivault, Olivier Dameron, and Nolwenn Le Meur, *Ontologies biomédicales et web sémantique pour la réutilisation des bases de données médico-administratives en pharmaco-épidémiologie*, JFO 2018-7ème Journées Francophones sur les Ontologies, 2018, pp. 1–6.



Philippe Tuppin, J Rudant, P Constantinou, C Gastaldi-Ménager, A Rachas, L De Roquefeuil, G Maura, H Caillol, A Tajahmady, J Coste, et al., *Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'assurance maladie (sniram) to the système national des données de santé (snds) in france*, Revue d'epidemiologie et de sante publique 65 (2017), S149–S167.

References II



Marie Zins, Marc Cuggia, and Marcel Goldberg, *Health data in france: Abundant but complex*, *Medicine sciences: M/S* 37 (2021), no. 2, 179–184.