Insee
Mesurer pour comprendre

Paris Brain Institute
ICM

sndsTools

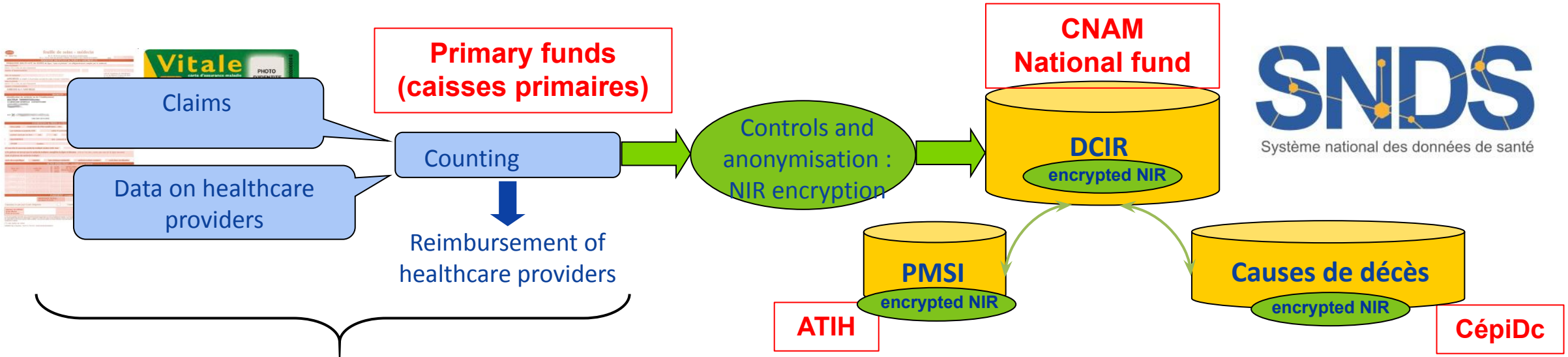
An R package to simplify SNDS data extraction

Matthieu Doutreligne
Insee
2026-01-26

Some slides from Antoine Belloir and Marc Dibbling

SNDS sources

January 2016 : created by law « modernisation de notre système de santé »



Local nominative databases

Reimbursement, possible identification, specific to each regime (three main regimes) and local fund.

Système National des Données de Santé (SNDS)

- **Anonymous national** data warehouse
- All regimes
- Chained with hospital data (**PMSI**, programme de médicalisation des systèmes d'information)
- Chained with **death certificates** (CépiDc)

🚀 **2021:** Permanent **access to research organisms** (INRIA, INSERM, hospitals…) : profil 108 and formation

In summary, SNDS is a rich database…

Populational data base

✔ ≃ 68 millions individuals

…but a claim database

Information not available

- ✖ Clinical exam results
- ✖ Basic clinical data (BMI, blood pressure, …)
- ✖ Environmental exposures or risk factors (smoking, drinking, nutrition habits)
- ✖ Socioeconomic status (education, income, …)
- ✖ Clinical data for hospital stays (beyond diagnostic codes) : eg. cheap medication during stays, biology, …
- ✖ No reason for consultation or prescription not reimbursed

In-town healthcare data (DCIR)

- ✔ Consultations
- ✔ Medical acts
- ✔ Biological exams
- ✔ Drugs and medical devices dispenses (reimbursed ones)

Hospital care data (PMSI)

- ✔ Inpatient hospital stay dates
- ✔ Medical acts and diagnostics
- ✔ Outpatient visits (consultations externes)

Death causes

- ✔ Medical diagnostics mentioned on the death certificate

Patient data

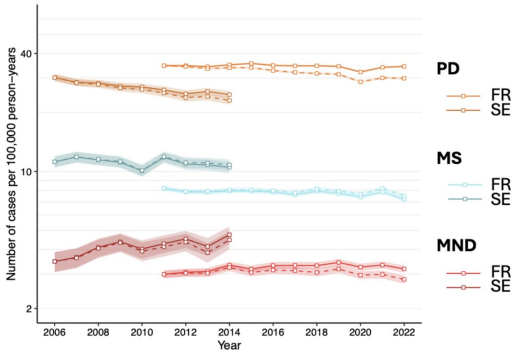
- ✔ Basic demographic variables (age, sex, department of residence, date of birth and death, geographical social deprivation index)
- ✔ Long Term Disease status with associated diagnostic codes
- ✔ Antecedents identified by the Cnam mapping system (“cartographie des pathologies”)

Overview of clinical research studies based on the SNDS database

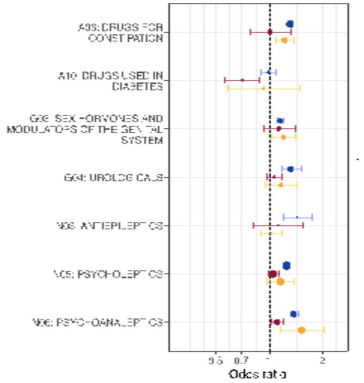
Epidemiological studies

Describe the health status of a population…

..and identify factors that influence it



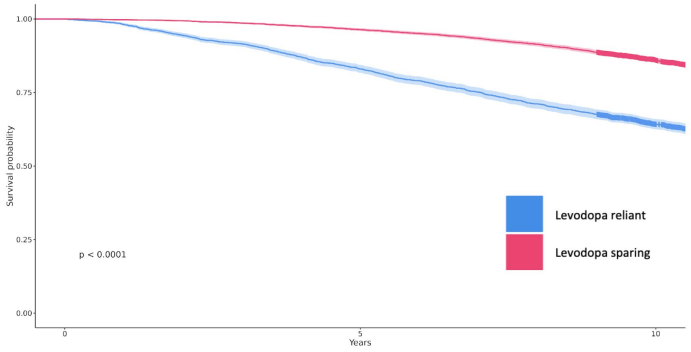
Incidence of Parkinson's disease in France over 10 years



Health conditions associated with Alzheimer's disease

Therapeutic studies

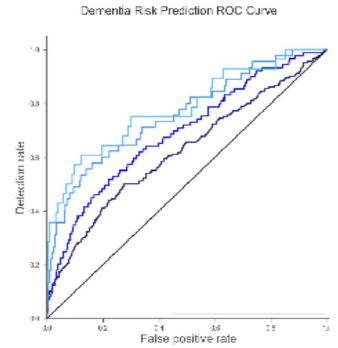
Compare the effectiveness of different treatments



Survival for two treatment options for 20 000 patients with Parkinson's disease

Prognostic studies

Predict disease progression based on individual patient characteristics

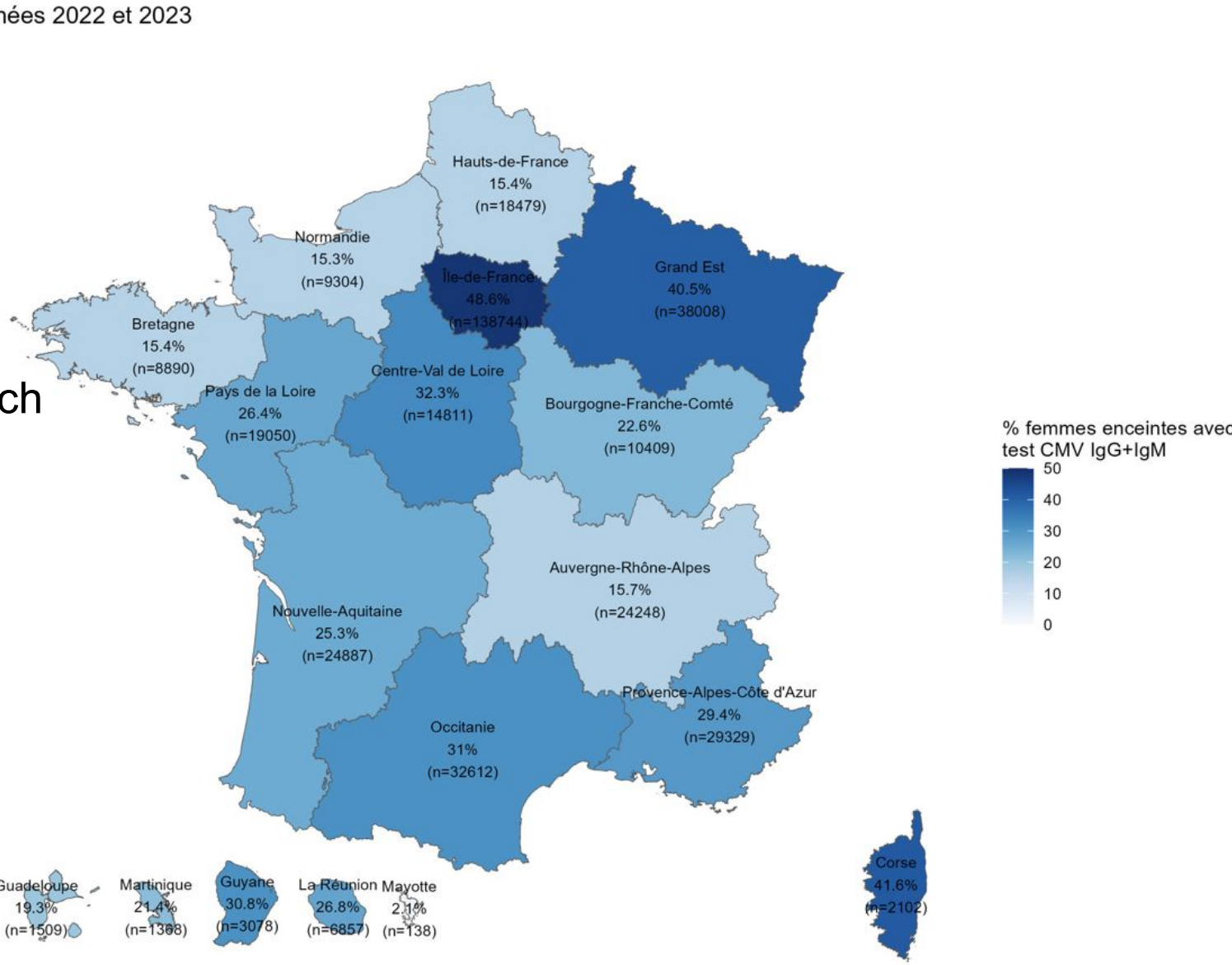


ROC curves for dementia onset prediction at 2, 5 and 10 years

Description of healthcare practices

Eg. Practice of screening pregnant women for CMV infection (2022-2023)

NABM 1260 : **IgG and IgM anti-CMV**

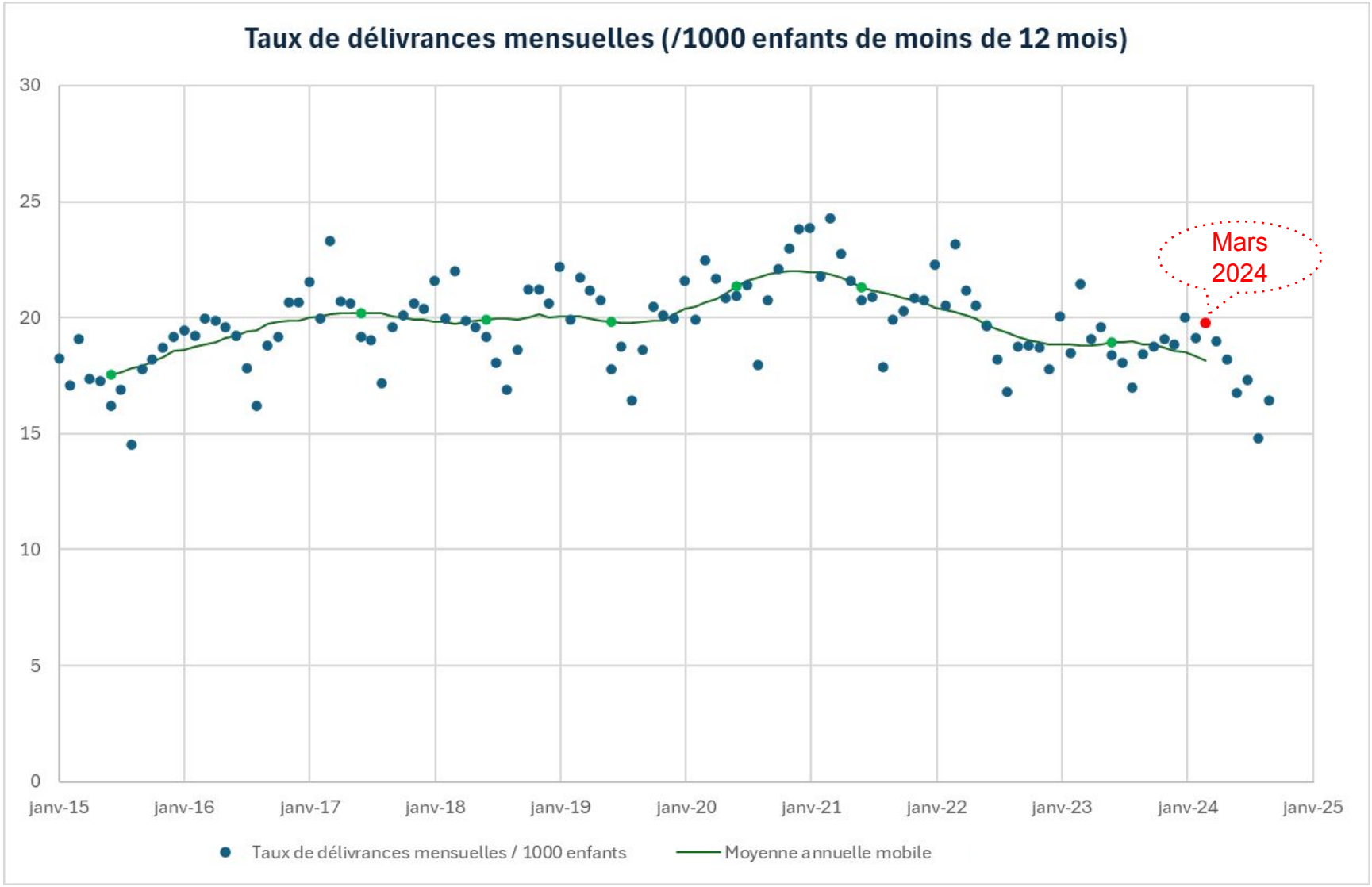


Heterogeneity of practices on the french territory

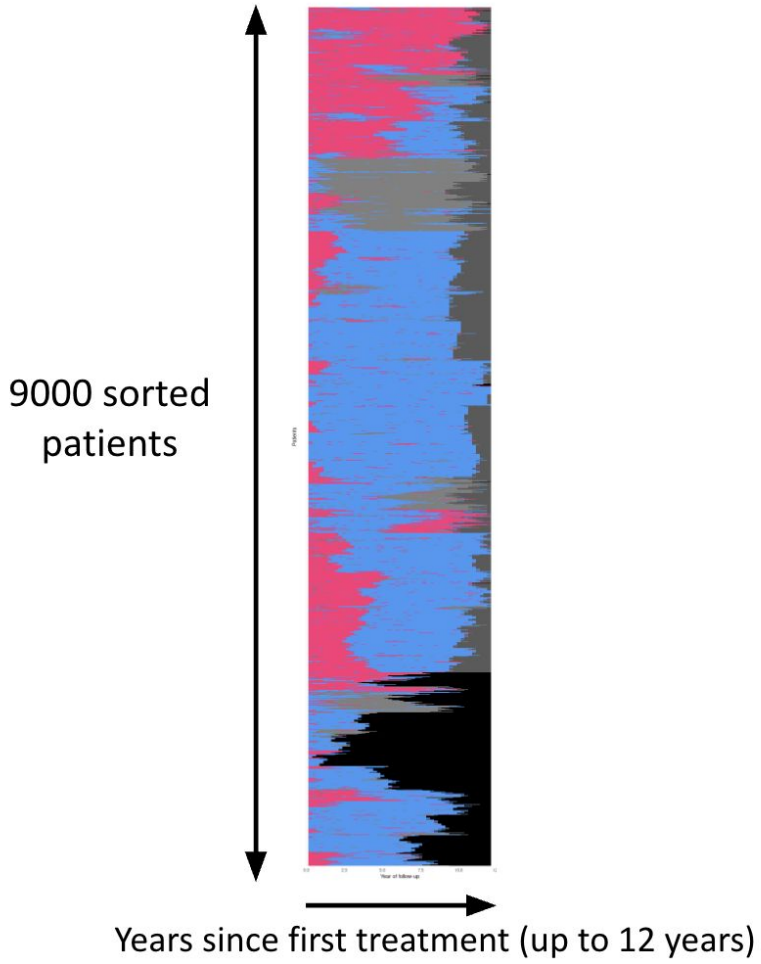
- Frequent use (> 40%) :
  - Île-de-France, Corse, Grand Est
- Much less common (≈ 15%) :
  - Bretagne, Normandie, Hauts-de-France, Auvergne-Rhône-Alpes

Impact of institutional recommendations

Eg. Proton pump inhibitor dispensing trends in children under 1 year of age since 2015 (⚠️ WIP)

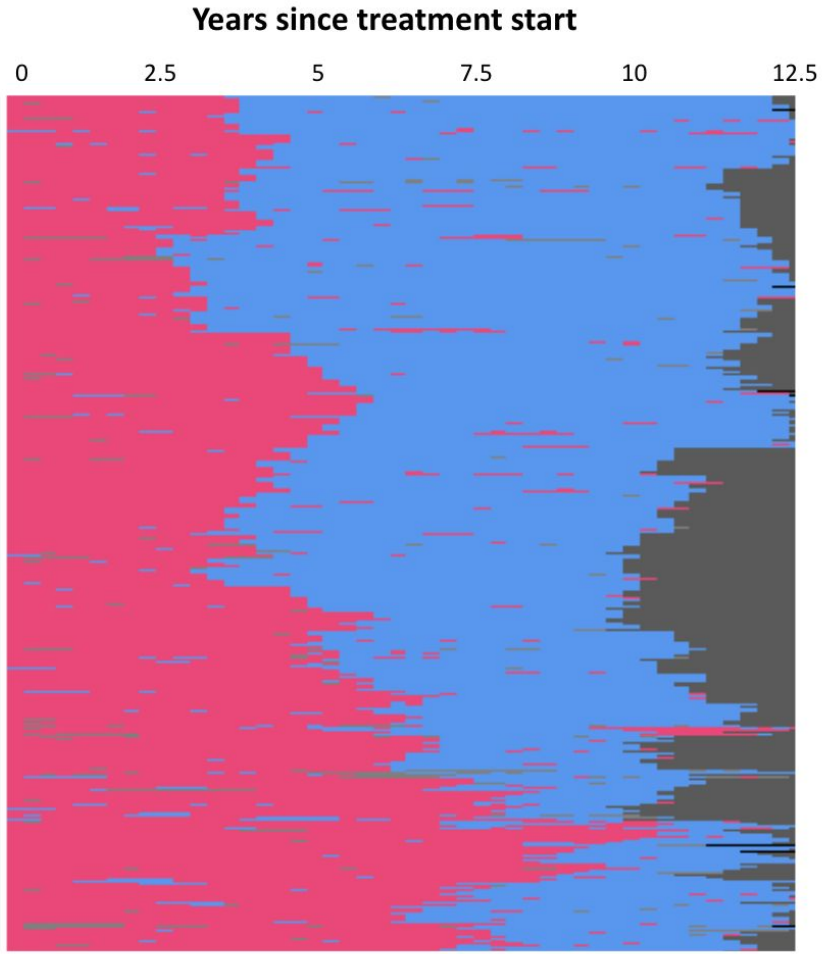


A sanity check in SNDS clinical studies : visualize individual treatment pathways

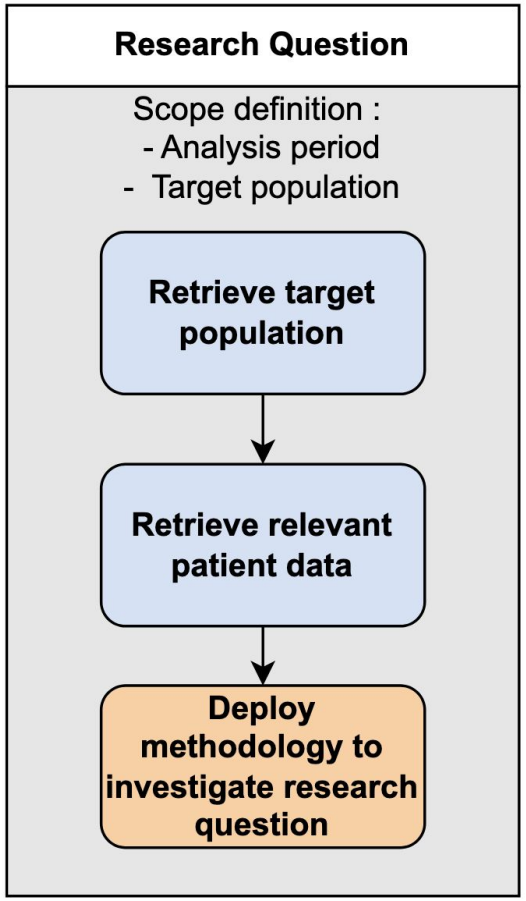


Patient treatment status during quarter

- Levodopa reliant
- Levodopa sparing
- No treatment dispensed
- Not yet observed
- Death



Typical workflow



Example :

Research Question
Heterogeneity in AD incidence over the last 10 years ?

Objective : Estimate the yearly incidence rate of AD patients and compare patient characteristics.

- **Analysis period** : 2013-2023
- **Target population criteria** : ≥ 2 AD codes at hospital OR ≥ 1 AD LTCD OR ≥ 3 dispenses of N06DA coded-drugs

Retrieve target population :
- Join patient reference table with tables of interest
- Filter depending on analysis period
- Filter and assess inclusion criteria
- Retrieve patient_ID

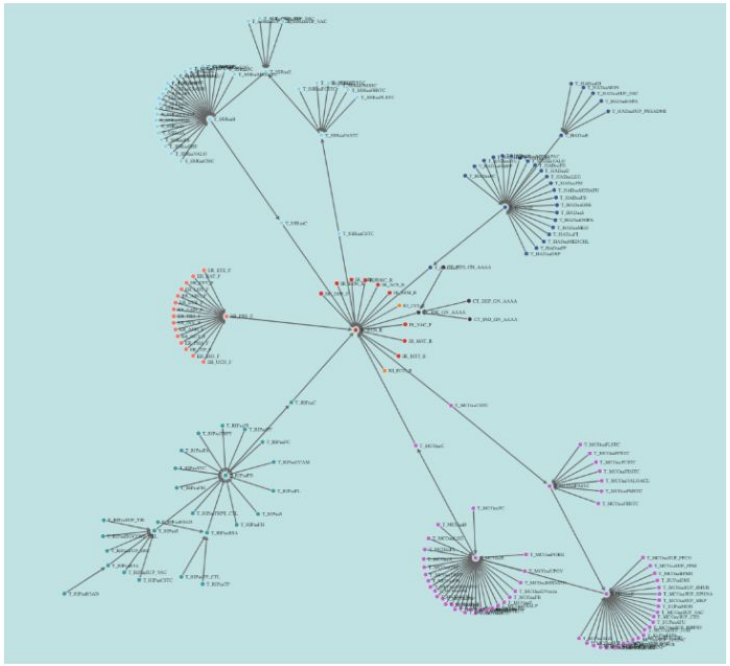
Retrieve relevant patient data :
- Retrieve patient characteristics & comorbidities
- Extract data from other tables relevant to the included patient_IDs

Deploy methodology to investigate research question
- Count yearly incidence cases
- Generalized Linear Model to explain heterogeneity

Typical workflow in practice

Data Richness : A Blessing and a Curse

Complex Data Structure



- Finding relevant variables
- Finding relevant tables
- Gathering, consolidating and reconciling data
- Looping over partially yearly indexed tables

Complex Data

- Tables containing (>200 columns and >100M lines)
- Filtering often required prior collection
- Managing duplicates, fictional data, errors, inconsistencies, date formats
- Specialized & Technical jargon

Technical hurdles

Limited data storage,
Limited R Package availability

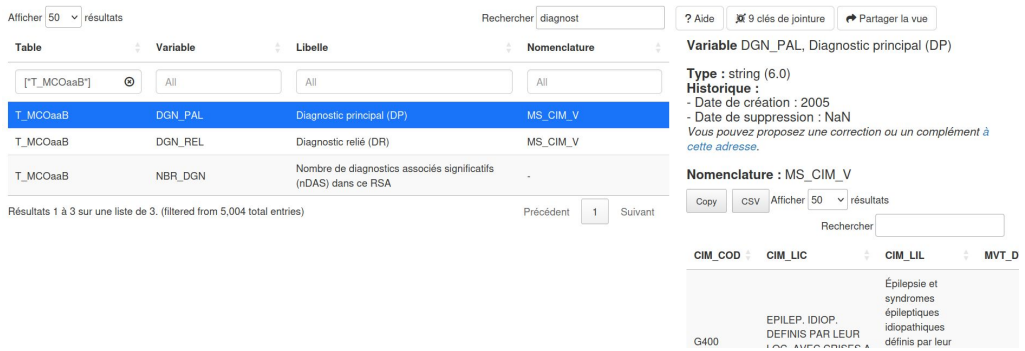
User's Guide:

428 pages long with excerpt code and crucial information to work correctly with SNDS data

Challenging for new users & many sources of errors (even for experienced ones...)

Resources for using the SNDS in clinical research

Interactive data dictionary :
https://health-data-hub.shinyapps.io/dico-snds/



A helpful collaborative documentation

Documentation collaborative du SNDS

Bienvenue sur la documentation collaborative du Système National des Données de Santé.

Cette documentation est en construction, via ce dépôt GitLab .

Contributeurs

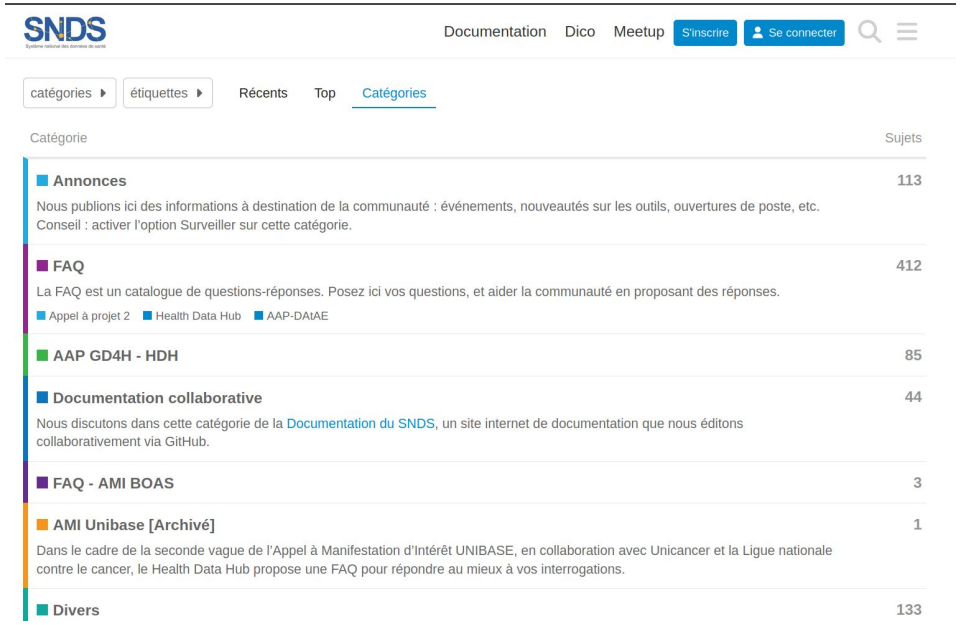
Cette documentation est maintenue par le Health Data Hub.

Elle résulte d'une mise en commun de documents et travaux par plusieurs organisations, dont :

- la Caisse nationale d'assurance maladie - Cnam
- le Health Data Hub - HDH
- Le Ministère des Solidarités et de la Santé: la Direction de la Recherche, des études, de l'évaluation et des statistiques - DREES
- les Agences régionales de santé - ARS
- Santé publique France - SpF
- la Direction de la Sécurité Sociale - DSS
- l'Agence Technique de l'Information sur l'Hospitalisation - ATIH
- le Centre d'épidémiologie sur les causes médicales de décès - CépiDC
- l'Institut national de la santé et de la recherche médicale - Inserm
- la Haute Autorité de Santé - HAS

https://documentation-snds.health-data-hub.fr/snds/

A forum



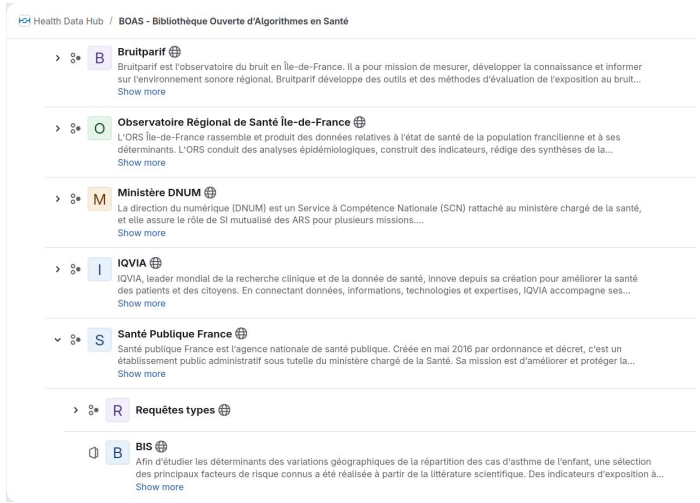
https://entraide.health-data-hub.fr/categories

Resources for coding with the SNDS

Several institutions use the SNDS database “at scale”, leveraging large *internal code bases* built over the years that are **not public**.

HDH BOAS: useful references of studies

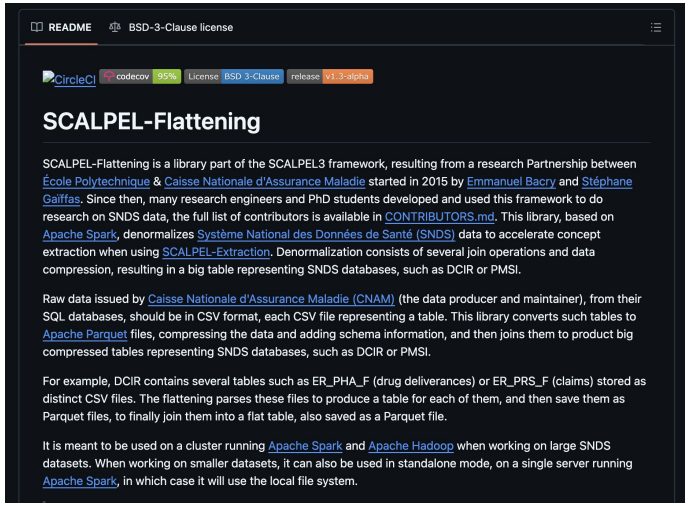
Mainly ad-hoc SAS code hard to adapt or configure



https://gitlab.com/healthdatahub/boas/

Open initiatives for a shared code base

Not adapted to shared infra and hard for collaboration

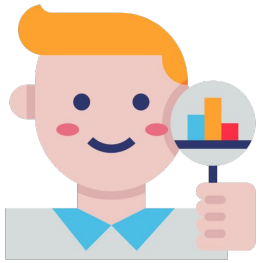


https://github.com/X-DataInitiative/SCALPEL-Flattening

Resources for using the SNDS in clinical research

To this day : no comprehensive code base for “new SNDS actors” to leverage the SNDS database with the existing CNAM platform

Consequences for individual researchers



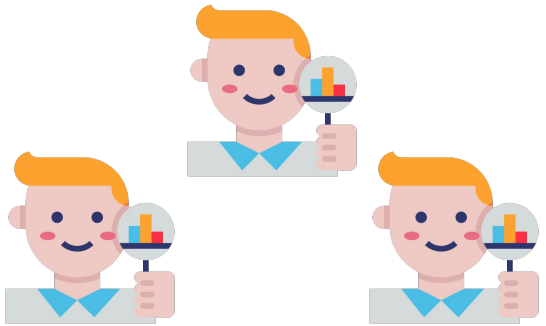
- Re-invent the wheel by re-implementing basic functions
- Re-implementation is typically :
  - compromise between quality and time
  - under-optimized
  - prone to errors or approximations

Time spent building a dataset >> time left for data analysis

Resources for using the SNDS in clinical research

To this day : no comprehensive code base for “new SNDS actors” to leverage the SNDS database with the existing CNAM platform

Consequences for research labs, public sector and research community

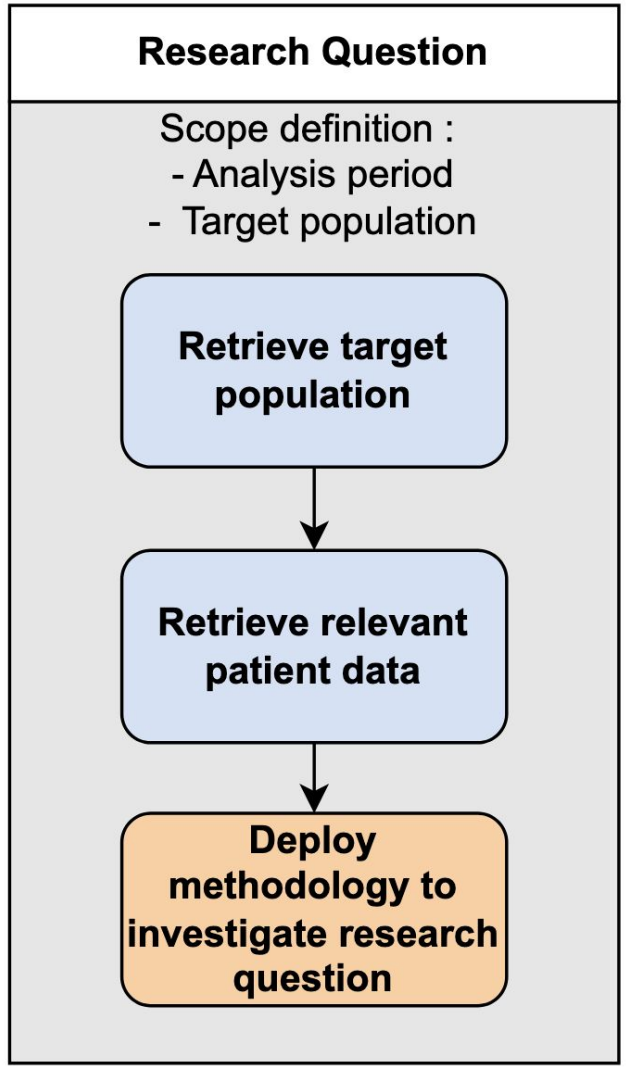


- Unnecessary “barriers to entry” ⇒ less research actors
- Duplicated work ⇒ waste of time and money
- Lack of uniformity and reproducibility when building SNDS datasets ⇒ improvable research quality

The sndsTools R package

Objectives

- 🔁 Enhance reproducibility across studies
- 📄 Improve readability
- 🤗 Facilitate onboarding for new users
- 🚀 Accelerate population extraction for all
- 🤝 Encourage code sharing and transparency
- 🌐 Community-driven



What is there right now in sndsTools?

Function references

https://sndstoolers.github.io/sndsTools/reference/index.html

✅ All functions are unit tested

🚧 Ongoing

🧐 Waiting review

PR69 - Example study

⏳ In review

PR64 - Extract MCO consultations with CCAM codes

Package index

Extraction SNDS

Fonctions pour extraire les données de soins individuelles à partir du SNDS.

`extract_consultations_erprsf()`
Extraction des consultations dans le DCIR.

`extract_drug_dispenses()`
Extraction des délivrances de médicaments.

`extract_hospital_consultations()`
Extraction des consultations externes à l'hôpital (MCO).

`extract_hospital_stays()`
Extraction des diagnostics des séjours hospitaliers (MCO).

`extract_long_term_disease()`
Extraction des Affections Longue Durée (ALD)

📚 Strong focus on documentation

- **Consistent API for extraction functions**
- **Documentation of :**
  - **Parameters**
  - **Outputs**
  - **Filter or SNDS choices**
  - **Link to original sources of documentation if existing**
- **Using roxygen2 to have the doc and the code both in the same R file.**

Extraction des consultations dans le DCIR.

Source: `R/extract_consultations_erprsf.R`

Cette fonction permet d'extraire les consultations dans le DCIR. Les consultations dont les dates `EXE_SOI_DTD` sont comprises entre `start_date` et `end_date` (incluses) sont extraites.

Usage

```
extract_consultations_erprsf(
  start_date,
  end_date,
  pse_spe_filter = NULL,
  prestation_filter = NULL,
  dis_dtd_lag_months = 6,
  patients_ids_filter = NULL,
  output_table_name = NULL,
  conn = NULL
)
```

Arguments

start_date

Date. La date de début de la période des consultations à extraire.

On this page

- Usage
- Arguments
- Value
- Details
- Examples

https://sndstoolers.github.io/sndsTools/reference/extract_consultations_erprsf.html

A simple case study as a tutorial

- **Fictive case :**
  hospitalized AVC in MCO
- **Demo of extraction functions**
- **Same code running :**
  - **Inside the CNAM server**
  - **Outside with fictive data**

Consistent documentation and code !

Not merged yet
ideas for improvement ?

```r
# Définir la période d'étude - année 2024
start_date <- as.Date("2024-01-01")
end_date <- as.Date("2024-12-31")

# Codes CIM-10 pour les AVC
codes_avc <- c("I61", "I62", "I63", "I64")

# Extraire les séjours avec diagnostics d'AVC
extract_hospital_stays(
  start_date = start_date,
  end_date = end_date,
  dp_cim10_codes_filter = codes_avc,
  or_dr_with_same_codes_filter = TRUE,  # Inclure les diagnostics reliés
  or_da_with_same_codes_filter = FALSE,  # Exclure diagnostics associés similaire
  and_da_with_other_codes_filter = FALSE,  # Exclure diagnostics associés différe
  da_cim10_codes_filter = NULL,  # Pas de filtre sur diagnostics associés
  patients_ids_filter = NULL,  # Extraire tous les patients
  output_table_name = "TMP_SEJOURS_AVC",  # Stocker en table Oracle
  conn = conn
)
#> Results saved to table TMP_SEJOURS_AVC in Oracle.
#> NULL

# Récupérer un aperçu des données
sejours_avc_head <- dplyr::tbl(conn, "TMP_SEJOURS_AVC") |>
  head(5) |>
  dplyr::collect()

kable(sejours_avc_head)
```

ETA_NUM	RSA_NUM	SEJ_NUM	SEJ_NBJ	NBR_DGN	NBR_RUM	NBR_ACT	ENT_MOI
190076	3	3	10	3	2	6	7

https://sndstoolers.github.io/sndsTools/pr-preview/pr-69/articles/tutoriel_avc.html

📌 Next steps ?

- Extraction of top pathologies
- Extension to other PMSI fields : HAD, SMR
- Extension to other SNDS fields :
  sick leaves, expensive drugs in ER_UCD_F, …
- 🤗 Search for contributors :
  https://github.com/SNDStoolers/sndsTools/issues
  - Ideas
  - Review
  - Code

Différente manière de contribuer

Répondre à une question sur une issue

De nombreuses questions sont posées sur les issues. Vous pouvez y répondre en donnant des conseils, ou en proposant une solution.

Créer une nouvelle issue

Si vous avez trouvé un bug, ou si vous avez une question sur une fonctionnalité, vous pouvez créer une nouvelle issue.

En cas de bug, il est important de donner un exemple reproductible `reprex` . Celui-ci contient le code nécessaire pour reproduire le bug, et le message d'erreur complet. Il est très important pour qu'un développeur plus expérimenté puisse comprendre le problème et vous aider.

Contribuer à la documentation

Contribuer à la documentation est aussi important que de contribuer au code. Vous pouvez proposer des modifications à la documentation en créant une pull request. Les petites erreurs et modifications peuvent être corrigée directment dans l'interface web GitHub.

NB: La documentation est principalement dans le code R, et est générée avec le paquet roxygen2. Pour la modifier il faut donc modifier les fichiers `.R` dans le dossier `R/` du projet.

Contribuer au code

Afin de résoudre un bug, ou d'ajouter une nouvelle fonctionnalité, vous pouvez créer une pull request.

Thanks to all contributors :

Antoine Belloir (ICM)
Thomas Soeiro (AP-HM)
Marc Dibbling (ICM)
Leo Guillon (ICM)
Catherine Bisquet (HAS)
Nelly Leguen (HAS)

Supplementary slides

An example of case study used as benchmark

https://sndstoolers.github.io/sndsTools/articles/benchmark_dplyr_vs_rsql.html

- Is there a speed difference between dplyr syntax and direct sql queries?
- Simple yet big query on ER_PRS_F, ER_PHA_F and ER_ETE_F
- Case study: C vitamin drugs on different period of times
- CCL: No big differences!