

Landscape of data sources for real-world observational studies: focus on the French context

OBJECTIVE

To provide a comprehensive overview of data sources and associated regulatory frameworks available for researchers, study sponsors and regulators in order to investigate the use, safety and effectiveness of therapeutic innovations in France.

METHODS

A structured literature review was conducted to identify existing data sources relevant for real-world observational studies. Both local and international initiatives were examined to ensure the completeness of the review. Giving the large number of data sources, we refined the scope of work to France.

RESULTS

The review identified four main categories of data sources which can be re-used for research

- Primary data collection including data collected for a specific study, by clinical research associates (CRAs) or digital tools such as PRO (patient queries)
- Reuse of reimbursement data including medico-administrative databases from hospital and primary care reimbursement
- Reuse of healthcare data collected during patient care or directly collected for or scientific research purposes, including access to hospital or primary care data warehouses.
- Finally, environmental and socio-economic data, publicly available on national websites or collected through sensors, can be linked to previous data sources using geographic identifiers to enrich analyses. We focused only on open data access as as the other data sources are too specific.

Regulatory requirements vary according to the type of data source but most of them involve obtaining authorization from the data owner and approval from ethics committee.

















CONCLUSION











This work provides practical guidance for researchers in selecting the most suitable data sources for real world studies and outlines the key regulatory processes.





































Glossary

- PMSI:** Medical Information Systems Program used in French healthcare institutions (hospitals, clinics, etc.) to record, code, and analyze medical and medico-technical activity.
- SNDS:** French National Health Data System including pseudonymized data from Health insurance data, Hospital data from the PMSI, Causes of death from civil registries, Disability and accidents data, Vaccination and preventive care data.
- DCIR:** Inter-Scheme Consumption Data containing individual-level data on all reimbursed healthcare expenditures in France, drawn from the different health insurance schemes.
- CépiDC:** Causes of death from civil registries.
- HUGO, APHP, CHU Nantes, CHU Reims:** Data from hospital data warehouses.
- BIOGROUP:** Cohorts of patients monitored in private practice laboratories.
- URO CCR:** National multicenter registry dedicated to the collection of clinical, biological, and radiological data on kidney cancer patients.
- Colibri:** National vaccination database.
- ATMO:** Federation of regional air quality monitoring associations, provides air pollution measurements.
- Météo France:** Meteorological and climate data (temperature, precipitation, extreme events, climate normals).
- Citepa:** Inventories of greenhouse gas and pollutant.

RESULTS

Overview of data sources and associated regulatory frameworks and access terms							
	Primary data collection	Reuse of data collected for reimbursement		Reuse of data collected in the context of care or research			Environmental and socio-economic data
	Primary data collection	Reuse of data collected for reimbursement (hospital)	Reuse of data collected for reimbursement (primary care)	Hospital data	Primary care data	Epidemiological registries	Open data
Data categories available							
 Examples in France (not exhaustive)	Adhoc	PMSI, SNDS, DCIR CépiDC		HUGO, APHP, CHU Nantes, CHU Reims...	BIOGROUP	URO CCR, COLIBRI	ATMO, Météo France, CITEPA
 Access and regulation	CNIL authorization, local ethics and scientific committee	CNIL authorization, protocol, CESREES / HDH validation		CNIL authorization, local ethics and scientific committee			Open access / no regulatory process
 Time frame	18-24 months	12-18 months		12-18 months			Open access : 0
 Data access costs	  	No access fees		 			Open access and no access fees
 Representativeness	Adhoc 	National		 Hospital / inter-hospital  According to locations of private laboratories	Adhoc 		National 

Data categories available for each source and linkage potential							
	Primary data collection	Reuse of data collected for reimbursement		Reuse of data collected in the context of care or research			Environmental and socio-economic data
	Primary data collection	Reuse of data collected for reimbursement (hospital)	Reuse of data collected for reimbursement (primary care)	Hospital data	Primary care data	Epidemiological registries	Open data
Data categories available							
 Socio-demographics	✓	✓	✓	✓	✓	✓	✓
 Patient pathways		✓	✓				
 Diagnostic and acts	✓	✓		✓	✓	✓	
 Medical test results	✓			✓	✓	✓	
 Drugs	✓	✓	✓	✓	✓	✓	
 Causes of death		✓	✓				
 Cost of stays		✓	✓				
 PRO, qualitative data from patients	✓	✓			✓	✓	
 Other							✓
 Linkage feasibility	+++			+++	++	+	

Pros & cons of using data sources							
	Primary data collection	Reuse of data collected for reimbursement		Reuse of data collected in the context of care or research			Environmental and socio-economic data
	Primary data collection	Reuse of data collected for reimbursement (hospital)	Reuse of data collected for reimbursement (primary care)	Hospital data	Primary care data	Epidemiological registries	Open data
Data categories available							
 On demand collected data							
 Confirmed diagnostic							
 Structured data							
 Unstructured data							
 Patient follow up							
 Level of authorities recognition							
Uses cases	<ul style="list-style-type: none">• Need for data generation• High level expectation of evidence for reimbursement, or drug authorization	<ul style="list-style-type: none">• Post drug authorization• Price negotiation				<ul style="list-style-type: none">• High level expectation of evidence for reimbursement, or drug authorization	<ul style="list-style-type: none">• Epidemiological studies, public health

Use cases examples				
	Primary data collection	Reimbursement databases	Patients records	Environmental and socio-economic data
Use-cases examples	Safety drug study	Patient pathway study	Evaluation of efficacy and treatment patterns	Public study on soil contamination
	