

CHALLENGE: accelerating research on cancer risk factors by structuring open data sources, and completing the OSIRIS clinical and -omics databases, in order to standardize variables related to the environment (terminology, interoperability,...)

OBJECTIVE: easing analyses of environmental cancer risk factor data by structuring and harmonizing open source epidemiological data sets, in a FAIR approach

WORK PACKAGES:

WP1: ontology
WP2: data & metadata
WP3: data sources

TARGET OUTCOME:

Standardized cancer epidemiology dataset framework & examples

NEOS FRAMEWORK (SELECTED FIELDS):

Current address • For how long • Past addresses (starting with most recent, as detailed as possible) • For how long (years) for each past address • Main occupation • Usual place of main occupation • For how long (years) • Main mode of transportation • How many days a month • How many hours a week

Consent (if needed)

Item group • Objectives • Item N° • Collection status • Item • Item definition • Expected value

Geographic location of measure • Geographic granularity of measure • Date of measure • Temporal granularity of measure • Data source • Geographic and temporal relevance

Main cancer sites associated with agent • Reference value • Guidelines • Monograph/backup paper • Main sources of exposure

Exposure to carcinogen (concentration in medium)

SCOPE:

YES

Cancer

Risk factors

Environmental

Air & water

IARC Group 1

Open sources

France

Epidemiology

NO

Other diseases

Prognostic factors

Behavioral & occupational

Other agents

IARC groups 2A, 2B, 3

Private sources

Rest of world

Care

RISK FACTOR SELECTION:

- Easily measurable → **air & water agents, France**
- International reference → **IARC** (International Agency for Research on Cancer) **monographs**
- Scientifically validated → **Group 1 carcinogens** (substances known to have carcinogenic potential for humans)



EXAMPLES:

From a list of 37 IARC Group I air & water biological, chemical and physical agents with open source data, we selected two carcinogens:

- An air pollutant: **PM 2.5** (fine particle matter), associated with lung cancer risk
- A water pollutant: **arsenic**, associated with lung, urinary bladder and skin cancer risk

CONCLUSIONS:

- Open source environmental data are very heterogeneous
- Two types of data are crucial for the NEOS Framework: place of residence/occupation, total duration of exposure.
- Definition of variables must be in context and precise to avoid bias
- Data collection and analyses at the patient level require a precise address and geocoding.
- This work will be expanded to other IARC Group I environmental cancer risk factors with open sources, using the NEOS Framework

Possible limitations, particularly for rural areas, include the place where measurements are obtained, and agents' geographical coverage.

