NEOS

NEoplasm Open-Source Environmental Risk Factors Standardization

Challenge 2

Pascal Deschaseaux[1], MD, MBA; Sébastien de Longeaux[1], MBA; Edouard Debonneuil[2], PhD; Rachel Aronoff[3], PhD

Air agents | Water agents

CHALLENGE: accelerating the research of cancer risk factors by structuring open data completing the OSIRIS clinical and omics data structure in order to include variables related to the environment (structuring, terminology, interoperability,..)

OBJECTIVE: easing environmental cancer risk factors data interoperability, exchange and further analysis by structuring, harmonizing an sourcing a minimal epidemiological data set, in a FAIR approach

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

TARGET OUTCOME:
Standardized air & water epidemiology dataset framework & examples

SCOPE:

Other diseases	Prognostic factors	Behavioral & occupational	Other agents	IARC groups 2A, 2B, 3	Private sources	Rest of world	Care
Cancer	**Risk factors**	**Environmental**	**Air & water**	**IARC Group 1**	**Open sources**	**France**	**Epidemiology**

RISK FACTORS SELECTION:

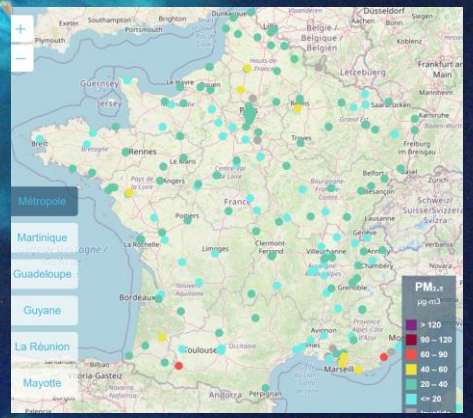
- Easily measurable → air & water agents, France
- International reference: IARC (International Agency for Research on Cancer, Lyon, France) monographs
- Scientifically proven → IARC group 1 carcinogens (substances known to have carcinogenic potential for humans, classification is largely based on human evidence)

EXAMPLES CHOSEN:
From a resulting list of 37 Group 1 air & water biological, chemical and physical agents with open sources, we selected 2 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:

- Environmental data are very heterogeneous
- Two types of data are key: place of residence/occupation, duration of exposure
- Variables definition must be very precise and contextualized to avoid biases
- The place of agent measurement, and its geographical coverage are possible limiting factors, particularly for rural areas
- Data collection and analysis at the individual level require a precise address and geocoding
- This work must now be expanded to the other IARC Group 1 agents using the NEOS framework



Maximal PM2.5 on January 11, 2022 according to Geodair.fr
It is not obvious to interpolate PM 2.5 throughout France !

NEOS FRAMEWORK (SELECTED FIELDS):

- Item group, objectives, item N°, collection status, item, item definition, expected value
- Geographic location of measure, geographic granularity of measure, date of measure, data source, reliability of use
- Main cancer sites associated with agent, reference value, guidelines, monograph/backup paper, main sources of exposure
- Consent (if needed)
- 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
- Exposure to carcinogen (concentration in medium)

Complete framework is available on Github: https://github.com/Epidemium/GeoStatus/NEOS

1 NewClin, 2 International Longevity Alliance, 3 Hackuarium

Acknowledgements: we thank the organizers of the Epidemium challenge (Roche, the operational team) as well as the project mentors (Mehdi Benchoufi, Julien Guérin, Bastien Rance, Hector Countouris) and reviewers