NEOS

NEoplasm Open-Source Environmental Risk Factors Standardization

Challenge #2

« 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,..) »

CONTEXT

World Health Organization

Preventing disease through healthy environments (WHO)

CANCER

■ DALYs due to preventable environmental risks (DALYs: disability-adjusted life years)
■ Proportion of disease attributable to the environment
Main areas of environmental action to prevent disease

OSIRIS

Proposed a minimal set of clinical and genomic items using international standards and terminologies enabling a strong interoperability. OSIRIS common data model is modular and extensible to other types of data

› JCO Clin Cancer Inform. 2021 Mar;5:256-265. doi: 10.1200/CCI.20.00094.

OSIRIS: A Minimum Data Set for Data Sharing and Interoperability in Oncology

Julien Guérin [1], Yec'han Laizet [2 3], Vincent Le Texier [4], Laetitia Chanas [1 5 6], Bastien Rance [7 8], Florence Koeppel [9], François Lion [10], Sophie Gourgou [11], Anne-Laure Martin [12], Manuel Tejeda [13], Maud Toulmonde [14], Stéphanie Cox [15], Elisabeth Hess [16], Marina Rousseau-Tsangaris [15], Vianney Jouhet [17 18], Pierre Saintigny [15 19 20]

Review › Environ Res. 2021 Jun;197:111185. doi: 10.1016/j.envres.2021.111185. Epub 2021 Apr 24.

Semantic standards of external exposome data

Hansi Zhang [1], Hui Hu [2], Matthew Diller [1], William R Hogan [1], Mattia Prosperi [3], Yi Guo [4], Jiang Bian [5]

« ... the heterogeneity of the external exposome data sources... increases the difficulty of analyzing and understanding the associations between environmental exposures and health outcomes. To solve the issue, the development of semantic standards using an ontology-driven approach is inevitable... »

THE TEAM: skilled (medicine, biology, maths, data and computational science, project management...), experienced (industry and public lifescience and IT R&D, open science, strategy...) and motivated!

Pascal Deschaseaux,MD, MBA

newclin
CONTEXTUALIZING HEALTH

Sébastien de Longeaux, MBA

newclin
CONTEXTUALIZING HEALTH

Edouard Debonneuil, PhD

INTERNATIONAL
·LONGEVITY·
ALLIANCE

Rachel Aronoff, PhD

Hackuarium

AGIR!

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

SCOPE

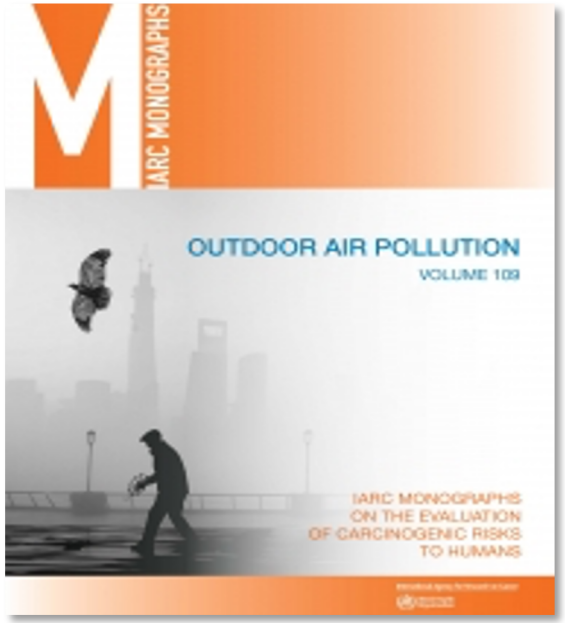
YES	**Cancer**	**Risk factors**	**Environmental**	**Air & water**	**IARC Group 1**	**Open sources**	**France**	**Epidemiology**
NO	Other diseases	Prognostic factors	Behavioral & occupational	Other agents	IARC groups 2A, 2B, 3	Private sources	Rest of world	Care

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)

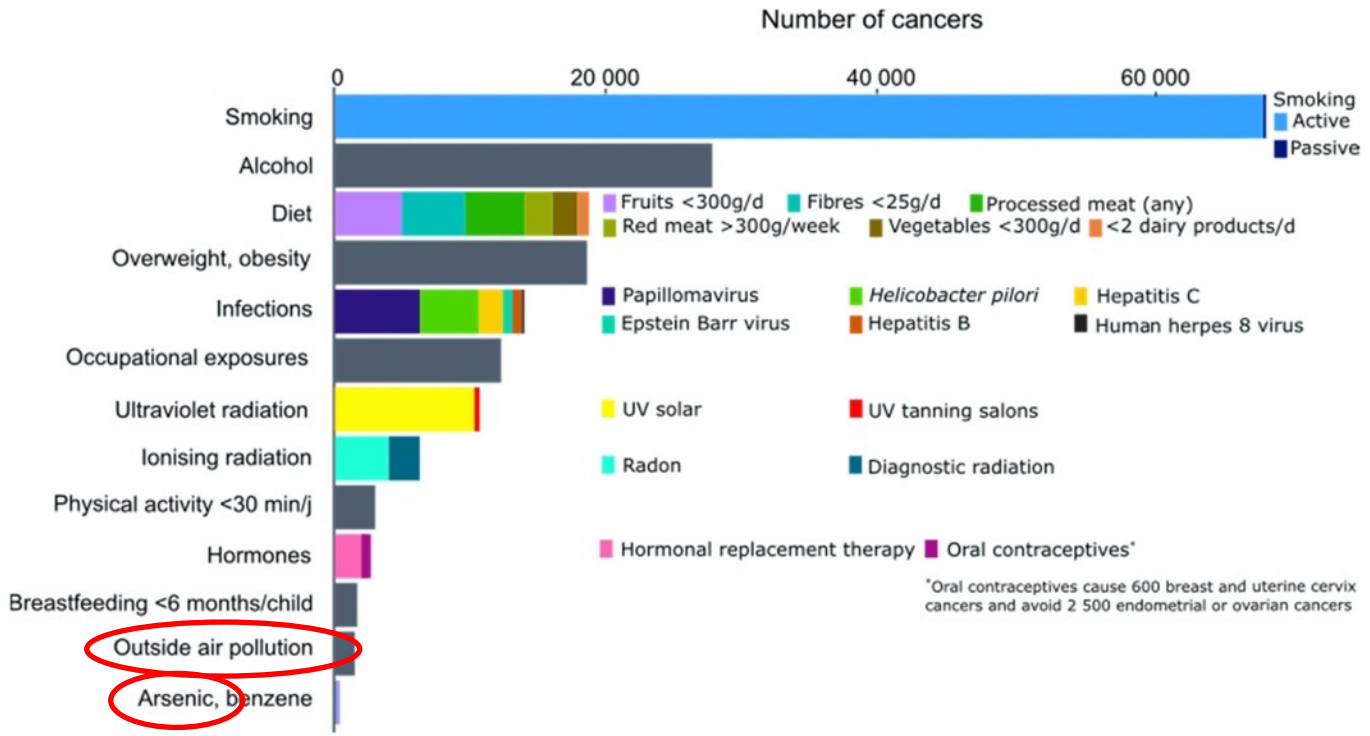


EXAMPLES

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

Numbers and proportions of cancer cases attributable to lifestyle and environmental factors in France in 2015, both sexes combined



I. Soerjomataram et al. / European Journal of Cancer 105 (2018) 103e113

DATA SOURCES & TOOLS

- IARC monographs
- Multiple open source databases on PM2.5 and arsenic, including studies of France's Institut de Veille Sanitaire
- France's Ministry of Environment beta web portal on health environmental risk factors
- Numerous bibliographical references
- Queries to OSIRIS and GEOCANCER specialists, Marc Fournier, Mehdi Benchoufi

- Collaborative tools
  - GitHub
  - JOGL
  - Slack
  - Drive
  - Inclusion of the NEOS project in OpenData4Health
- Google, Pubmed
- Google Image
- The OSIRIS datasets

NEOS FRAMEWORK (SELECTED FIELDS)

Available in an OSIRIS-like excel format

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

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

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

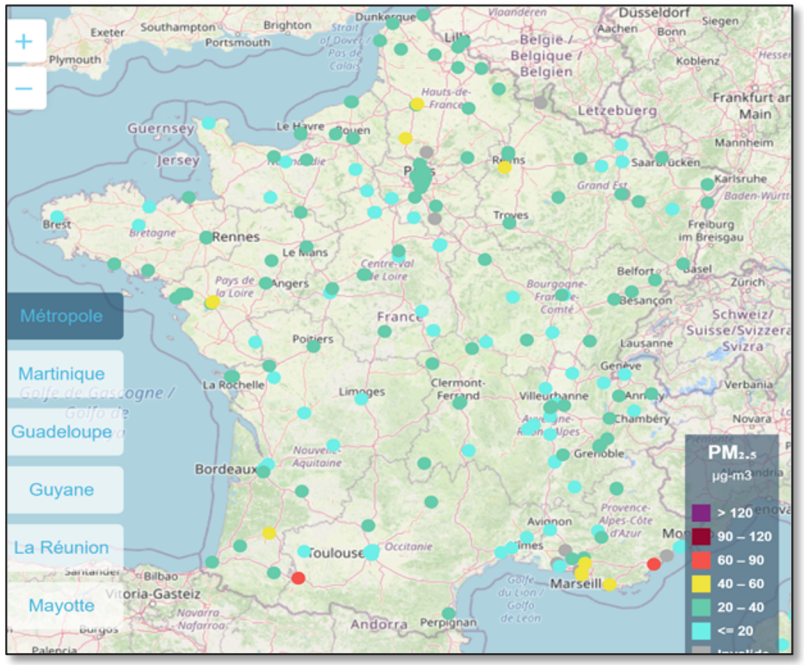
Exposure to carcinogen (concentration in medium)

Not agent-specific

CONCLUSIONS

- Open source environmental data are very heterogeneous: harmonization of references is highly needed for interoperability
- 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

Note: possible limitations, particularly for rural areas, include the place where agent measurements are obtained and their geographical coverage



Mean life expectancy gain at 30 years with the « no PM 2.5 atmospheric pollution » scenario (source : InVS, Santé Publique France). Map shows data measurement coverage, sometimes a single point per region

PERSPECTIVES

- In the coming months, this work will be expanded by our team and additional contributors to other IARC Group I environmental cancer risk factors with open sources, using the NEOS Framework

- Final use can be two-fold
  - **Etiology of cancers in diagnosed patients**
  - **Study of populations cancer risk for prevention**

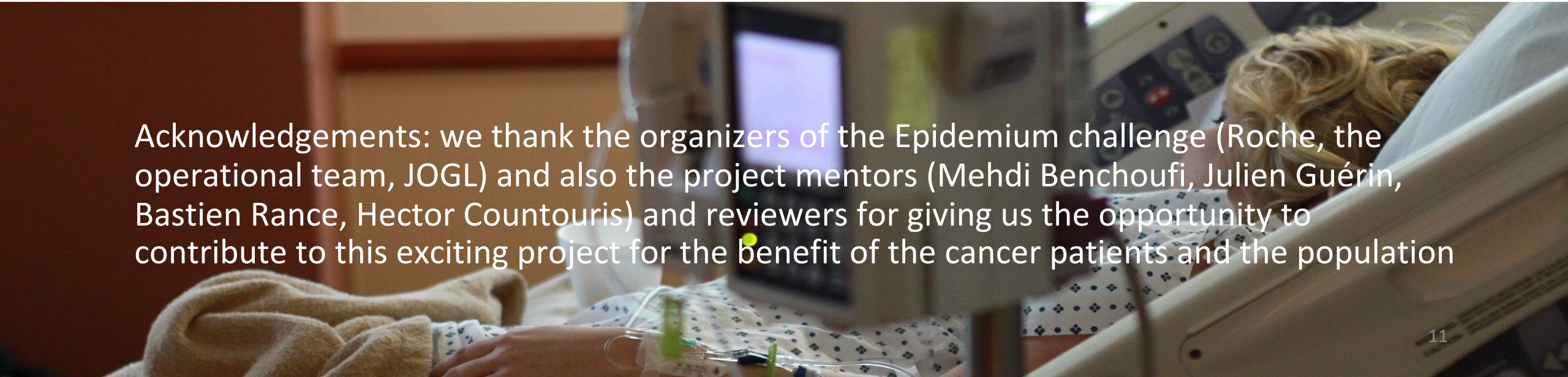
- Satellite and GIS data sources could help filling the data gaps in the future

Credit: Photo by Markus Spiske on Unsplash

THANK YOU!

See our documentation on:

https://github.com/Epidemium/EPIDEMIUM-Season-3/NEOS



Acknowledgements: we thank the organizers of the Epidemium challenge (Roche, the operational team, JOGL) and also the project mentors (Mehdi Benchoufi, Julien Guérin, Bastien Rance, Hector Countouris) and reviewers for giving us the opportunity to contribute to this exciting project for the benefit of the cancer patients and the population