H E L I C A L

H2020-MSCA-ITN-2018-813545

HELICAL

Health Data Linkage for Clinical Benefit

Deliverable D1.4

WP Summary Report

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 813545

HELICAL

Table of Contents

Work Package 1 Environmental impacts on autoimmunity

Develop methodologies to investigate the interaction between vasculitis onset and relapse
Lead: Mark Little (TCD)
Participants: P7D, UG, ISG, UNIABDN, KSG, Univ Padova, DFKI, IHD, TCD

AIM:

This WP is adopting a novel approach to the question of environmental impact on autoimmunity by studying the whole system in which the patient moves through time, comparing this to formal reductionist case-control study design. To achieve this, it is necessary to develop IT systems that facilitate fusing large quantities of data deriving from multiple sources, some of it sensitive; the first aim is to build on existing systems to create frameworks that support diverse dataset fusion and advanced analytics, linking environmental data streams with registry and app data. Focusing on the pre- and post-diagnosis periods in parallel, we are addressing the following questions: (a) Are there clusters within the population under study and can we identify specific factors associated with vasculitis flare and/or onset? (b) Can we predict for a given patient the probability of suffering a flare within a specified time? (c) Can an approach that combines agnostic machine learning and traditional focused case-control epidemiological study improve predictive fidelity? (d) Is it thus possible to develop a prototype physician/patient interface integrating environmental and patient level data to provide individual flare risk estimates, forming the basis for future clinical tool development?

EXPERIMENTS PERFORMED:

Enabling infrastructure

The twin challenges of studying environmental triggers of autoimmune disease are dynamic spatio-temporal linkage and irregular time series modelling using sparse datasets. These were addressed by ESRs 1 and 3.

ESR1 adopted a 3-phase approach:

1. Identify researcher requirements in linking data for environmental health research.
2. Develop a framework that enables a researcher to link environmental data with particular health events based on user data inputs.
3. Evaluate and refine the developed framework through rare disease case studies. This was undertaken iteratively through a series of systematic user interaction sessions involving 50 researchers.

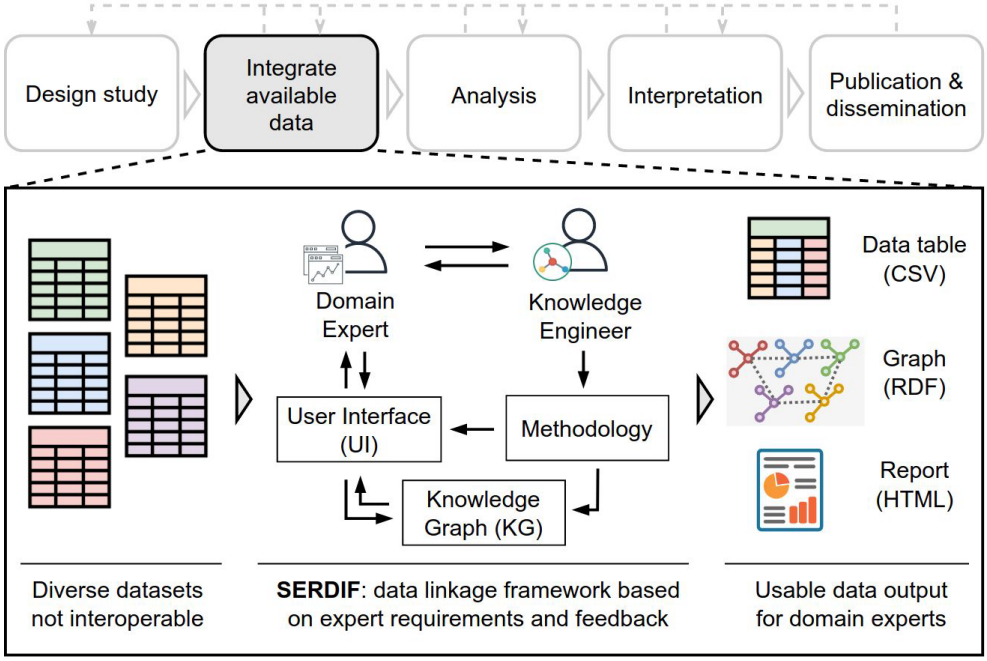


Figure 1. *Framework for development and evaluation of the SERDIF data integration tool.*

ESR3 addressed the challenge of developing a model that can detect a correlation between ANCA vasculitis flare propensity and environmental exposure. In brief, the approach comprised a specialised regression architecture, which accounts for the presence of a distributed lag period before relapse that can be inferred from the data in the context of an irregular time series. This integrates the Mixed-data sampling (MIDAS) model concept, generalising them for binary response data and progressing to a Bayesian implementation. This model performs variable selection and accounts for binary response imbalance. It was applied to linked environmental data derived from the Irish Rare Kidney Disease registry (using the software developed by ESR1), implemented in R and developed as an R package.

Linking registry data to environmental information

The two objectives of HELICAL WP1 are identification of triggers of vasculitis ***relapse*** and vasculitis ***onset***. By using the Irish RKD and UKIVAS registries, we studied **longitudinal** environmental triggers of relapse in patients *known to have AAV*. Thus, by considering the time series nature of each participant, it was possible to study spatio-temporal changes in environmental conditions and to link these to relapse at the subject level.

As patients who develop de novo AAV are generally not known before diagnosis, it was necessary for ESR4 to use large publicly available datasets to allow comparison to appropriate control groups: re-use of primary healthcare data from **NHS Scotland** and leveraging the **UK Biobank** to link new onset AAV with preceding environmental conditions.

The approach of ESR6 was focused on **time series analyses** to discover patterns within the temporal dynamics of disease incidences and establishing connections to environmental factors. Deep analysis of the spatiotemporal dynamics of Kawasaki Disease in Japan allowed ESR6 to test the hypothesis that

tropospheric winds play a role in the development of the disease. He conducted a comprehensive computational analysis of wind transport phenomena, as well as air sampling to assess the chemical and biological properties of the air at the destination. This has involved both the on-site air-sampling in Japan and the computational analysis of previous sampling campaigns.

Modelling approaches to exploring environmental triggers of autoimmune vasculitis

The relevant methodology is described in Deliverable 1.3. In addition to the MIDAS modelling approach described above, we used the following techniques:

1. The median prodrome period (that interval between onset of symptoms and recording of the clinically evident initial AAV diagnosis, or diagnosis of relapse) was estimated from existing registry data as being approximately 70 days. This derived value was applied in time series and data linkage analyses.
2. Relapse was studied using an **n-of-1 design** where each participant acts as their own control, eliminating confounding by time-independent factors such as gender, occupation and genotype. Case (relapse) and control (remission) windows were defined using an algorithm that considered the effect of residual disease activity (Figure 2). A multi-level model was applied to investigate the association between environmental factors (such as ambient vitD-UVB/CW-D-UVB) and AAV relapse. A random effect was included to account for repeated measures and the varying relapse risk between individuals.

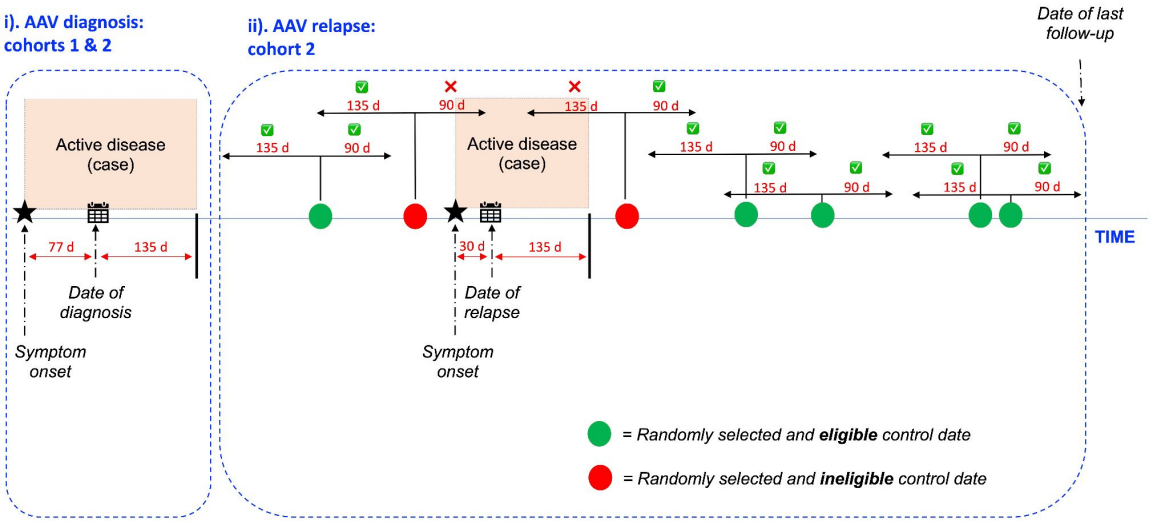


Figure 2. *Strategy for an n-of-1 study design to explore environmental triggers of vasculitis.*

3. To study AAV onset (diagnosis), logistic regression was used to examine the relationships between cumulative dose of pollution exposure and vasculitis risk, and between AAV phenotype and serotype (outcome variables) and measures of ambient vitD-UVB. Each of the 18-25 model's two-tail p-values were adjusted for multiple comparison using Simes-

Benjamini-Hochberg false discovery rate. A stratified analysis was undertaken to obtain insight into geographic variation of air pollution by population density.

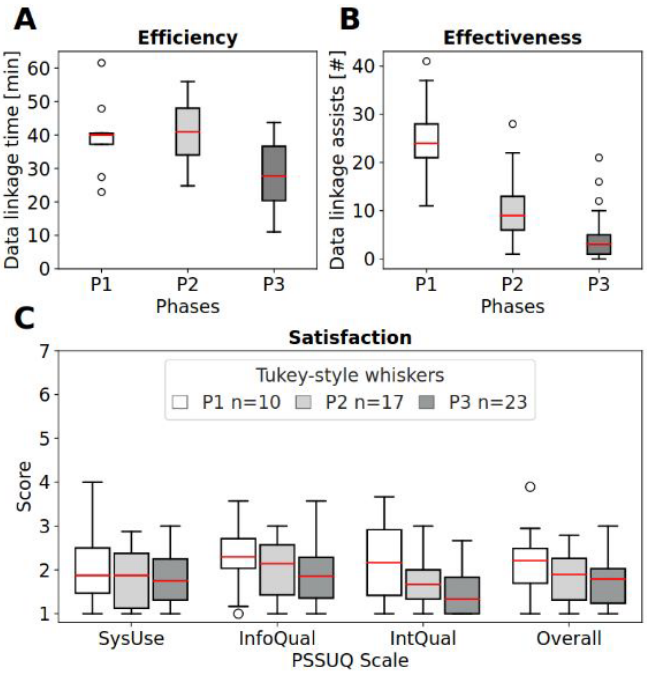
4. To evaluate the spatiotemporal dynamics in Kawasaki disease, we used a variety of methods.
    a. The implementation of Multiple Seasonal-trend Decomposition using LOESS (MSTL) served to break down the time series into trend, seasonal cycles, and residual components, offering an effective way to understand the underlying patterns in the data. In addition, to detect cycles of an uncertain periodicity that may be present in the series, Singular Spectrum Analysis (SSA) was employed.
    b. Hierarchical clustering, using the Ward method and Spearman correlation as a distance metric, was chosen for grouping regions based on the similarity of their temporal dynamics. For assessing similarities where the correlations between the time series might be transient and non-linear, Scale Dependent Correlation (SDC) analysis was applied.
    c. The task of evaluating the spatial autocorrelation of disease incidence was addressed with the use of Global Moran's I. To further supplement this analysis and determine the presence of spatial clusters, Local Indicators of Spatial Association (LISA) were used.

Case-control study

ESR5 designed a questionnaire-based study to compare lifetime occupational history between patients with AAV and community dwelling controls. The “Canjem” job exposure matrix was used to convert these occupations into a hierarchical list of specific pollutants. Latent class mixed models were used to study nonlinear associations between protracted exposures (e.g. asbestos and tobacco smoke) and vasculitis onset. Given extreme delay to commencement of this study, the same approach was applied to a publicly available Swedish dataset.

RESULTS:

A software tool for spatio-temporal environmental data integration



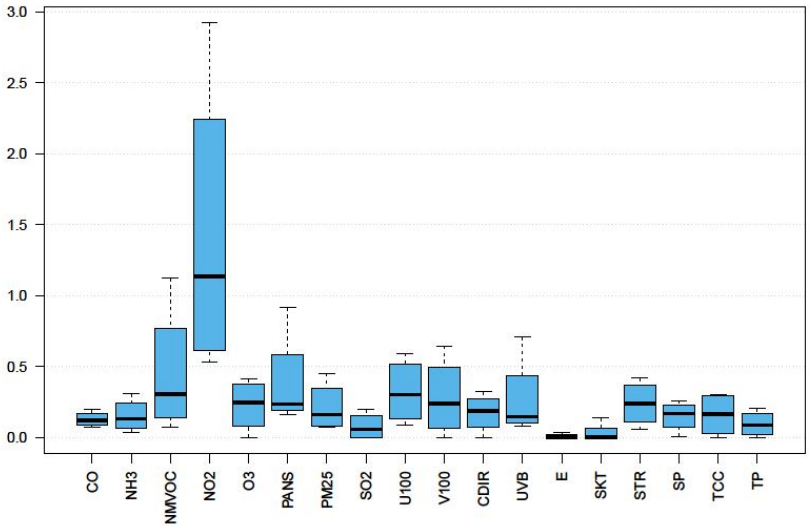
Extensive incremental evaluation of the SERDIF tool demonstrated progressive improvement in efficiency, effectiveness and satisfaction with each iteration (Figure 3). The tool is now applicable not only to the HELICAL use case, but more generically to any biomedical question that requires spatio-temporal linkage of patient location to environmental values.

Figure 3. *Summary of evaluation experiments illustrating progressive improvement in usability and satisfaction of the SERDIF tool.*

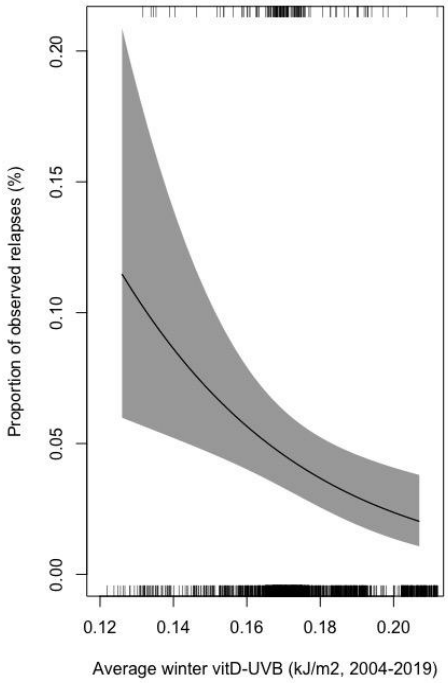
The association between environmental exposures and AAV relapse

The final dataset analysed using the MIDAS technique contained 87 patients from the RKD registry who had suffered a definite relapse, each with 4 years' worth of daily data and 18 location specific environmental explanatory variables. The Bayesian posterior prior values across all runs are given in Figure 4, which shows a low rate of variable selection. The most often selected value across all runs, at just under 3%, was Nitrogen Dioxide ($N0_2$). We see from the boxplot that $N0_2$ was consistently one of the most frequently selected variables, with the next most selected (on average) being Non-Methane Volatile Organic Compounds (NMVOC). No other variable achieved more than 1% inclusion across all 50,000 runs, and $N0_2$ was the only variable that never fell below 0.5%. These results suggest that, using this methodology and dataset, there is no clear association between the tested individual environmental triggers and vasculitis relapse, but $N0_2$ merits further investigation.

Figure 4. *Testing of association between environmental pollutants and relapse in occurrence in the Irish vasculitis population. Variable selection results after 4 runs with different starting points. Note the scale of the y-axis; across all runs, no variable is selected more than 3% of the time.*



The association between UVB exposure and AAV relapse and onset



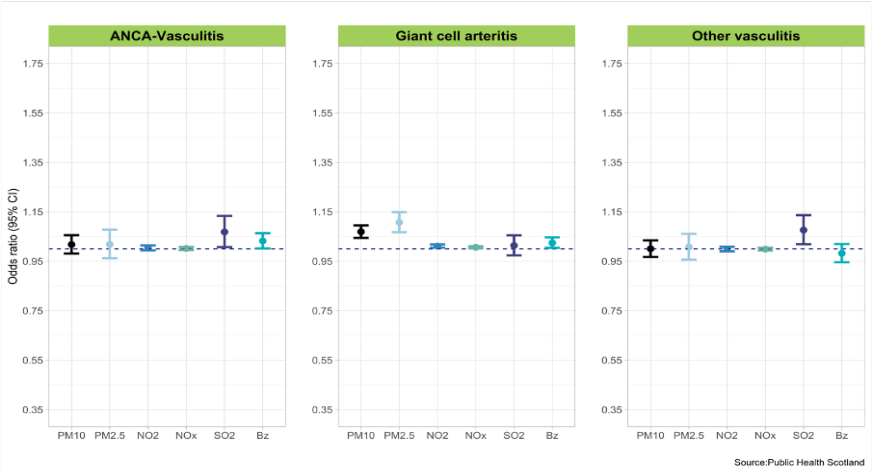
Residential latitude was positively correlated (OR:1.41, 95% CI 1.14-1.74, p=0.002) and average vitD-UVB negatively correlated (0.82, 0.70-0.99, p=0.04) with relapse risk, with a stronger effect when restricting to winter measurements (0.71, 0.57-0.89, p=0.002). However, contrary to our hypothesis, these associations were not restricted to granulomatous phenotypes. We observed no clear relationship between latitude, vitD-UVB or CW-D-UVB and AAV phenotype or serotype.

Figure 5. *Correlation between winter Vitamin D exposure and relapse.*

The association between AAV onset and air-borne pollutants

Analysis of the Scottish NHS dataset, with validation using UKIVAS and UK Biobank dataset illustrated a consistent association between sulphur dioxide, which was primarily observed in rural dwellers (Figure 6).

A



B

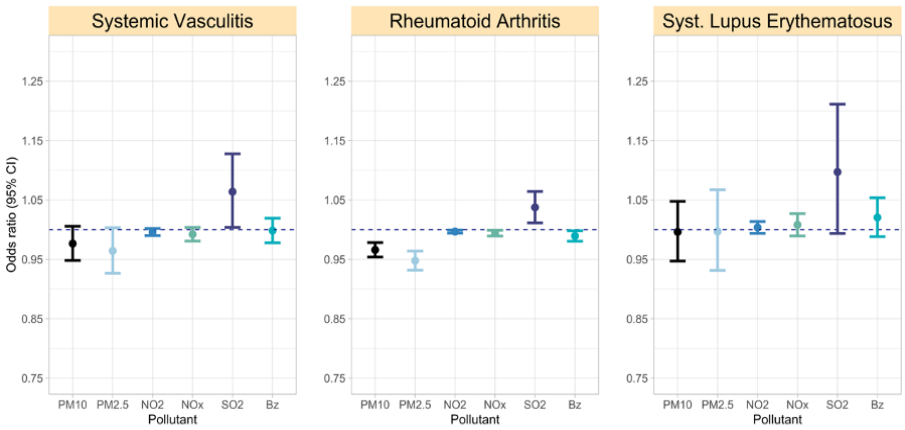


Figure 6. *Association between SO2 and vasculitis onset in Scotland (A), validated using UK Biobank data (B).*

Examination of prolonged occupational exposures on AAV occurrence

By applying a novel job exposure matrix to a publicly available AAV dataset, we observed associations with occupational exposure to 13 or 188 potential agents, including hydrogen sulphide, organic alkanes and aldehydes, and aromatic hydrocarbons, but not silica. These exposures tend to be enriched in agricultural and food processor workers (Figure 7). For example, 92% of substances associated with GPA are present in Agricultural and Animal Husbandry Workers, while only 17% are present in Cabinet Makers and related Woodworkers. These are often the product of acid catalysis of SO_2, providing a potential link between the data linkage and case control results. These findings are consistent with the concept of AAV as an oxidative disease.

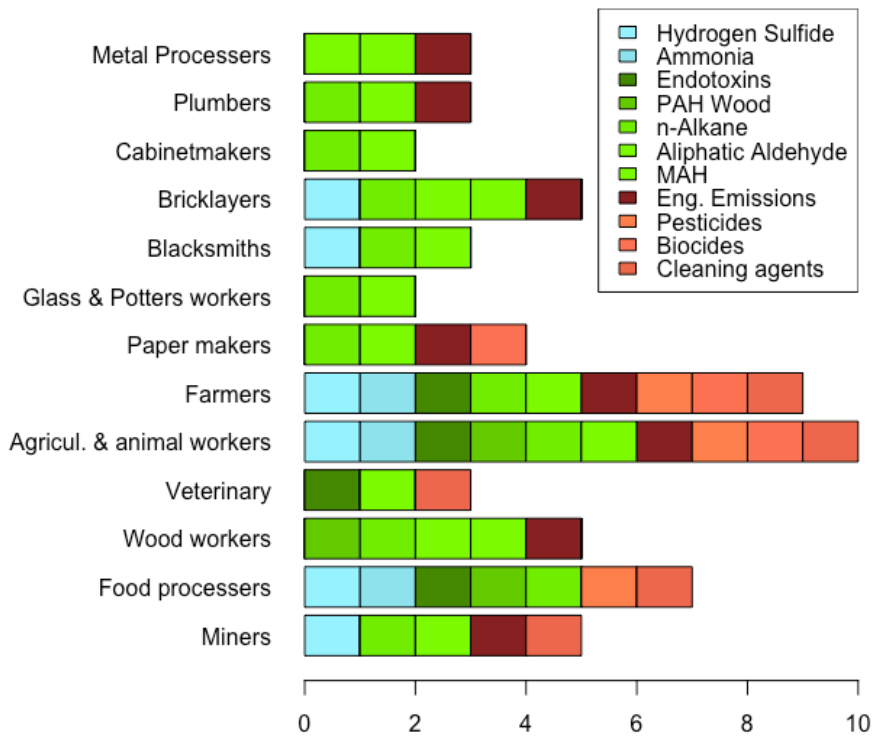


***Figure 7**. Distribution of the 11 substances found to be associated with GPA among the major 13 ISCO-68 groups. MAH = Monocyclic Aromatic Hydrocarbons*

Time series analysis of Kawasaki disease incidence

Daily variability of fine aerosols in a surveillance campaign in south Japan shows a striking co-evolution between their trace elements (metal and metalloid, MM) content and Kawasaki Disease (KD) admissions, suggesting a strong dynamical link. This association may account for >40% of total variability in the disease, being dominated by a clear sub-weekly cycle (SWC1). This SWC1 appears to connect or disconnect Japan to air intrusions from above the planetary boundary layer (PBL), having their source in industrial and agricultural areas in NE Asia, and points to a stronger case for an agricultural source for the exposure as opposed to urban pollution. KD maxima always occur in full synchrony with the arrival of very small (<1 µm; PM1) particles showing that ultrafine aerosols appear to be a necessary cofactor in the occurrence of KD (Figure 8).

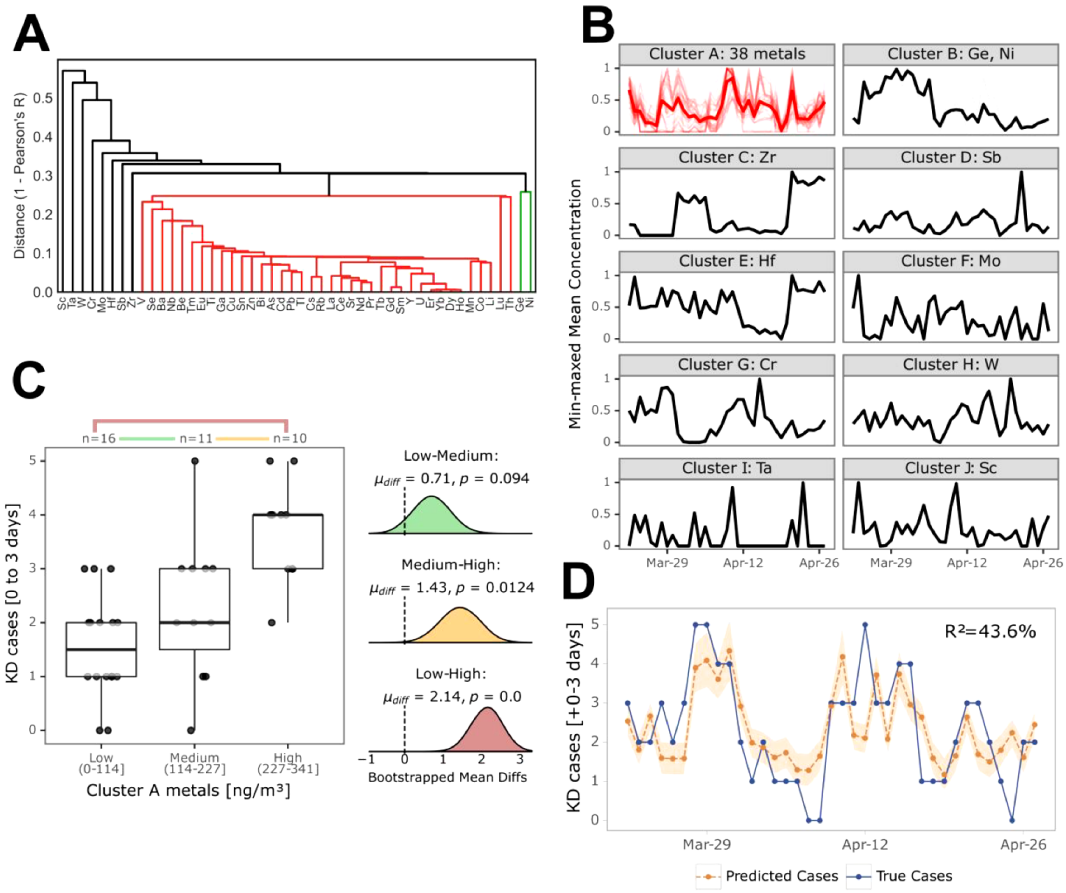


Figure 8. *Cluster analysis of the metal content in aerosols and its synchrony with KD incidence. (A) Dendrogram generated when using Pearson's correlation as distance metric and the Nearest Point Algorithm to compute distances across the newly formed clusters. Colours denote the different clusters assigned when setting a distance threshold of 0.3. (B) Variation during 37 sampled dates of the (min-max normalized) concentration of MM of each cluster. The thick line represents the median value of all cluster members at each time-point, with the individual metal contributions shown in the shaded thin lines. (C) Boxplots for the distribution of daily KD cases as a function of the daily concentration of Cluster A categorized into three groups: Low (0 to 114 ng/m³, n=16), Medium (114 to 227 ng/m³, n=11) and High (227 to 341 ng/m³, n=10). On the side, distribution of the difference of means for the pairwise comparison of groups performing a bootstrap test, showing significant differences in KD cases for days with high concentrations compared to those both in medium and low concentrations. (D) Reported (true) KD cases in the Kumamoto prefecture (blue) and predicted cases with a Poisson model using the concentration of Cluster A MM as predictor. The shaded area represents a bootstrapped 95% CI.*

KEY SCIENTIFIC OUTPUTS:

- SERDIF software tool (https://serdif-ui.adaptcentre.ie/) that uses knowledge graph technology to integrate patient location with environmental factors to allow study of the impact of these factors on disease progression. Critically, this software is open source and FAIR; its use of a flexible interoperable semantic web approach led to an unexpected scalability to other disorders. Indeed, the software is not limited to use in vasculitis. It can be applied in any condition where there is a need to link patient location over time with a wide range of environmental data streams, suggesting that it will have wider application in exposome research (Milestone 2: https://github.com/navarral/serdif/; described in publications 2, 12 and 14).
- This tool was used to study the association of Ultraviolet-B radiation and the occurrence and relapse of ANCA vasculitis, with the finding that relapse risk was associated with low cumulative winter UVB exposure. This work also demonstrated, for the first time, the use of an n-of-1 study design for investigation of longitudinal outcomes in patients (Publication 10).
- Development of an R-package for distributed lag modelling of unbalanced binary data through use of Bayesian quantile regression. Variable selection is possible and useful for quantifying potentialities of impacts of environmental exposures on binary responses (Deliverable 1.2; https://github.com/jsnwyse/dlvarsel).
- Identification of SO2 as a key environmental potentiator of ANCA vasculitis (Deliverable 1.3, manuscript in preparation).
- Description of strong consistent negative effects of both temperature and absolute humidity at large spatial scales associated with the spread of SARS-CoV-2 around the world. ESR6 classified, for the first time, COVID-19 as a seasonal low-temperature infectious disease (Publication 13).
- Development of a Python package, designed specifically for the execution of Scale Dependent Correlation (SDC) analysis. This tool facilitates the exploration of transient synchronicities within time-series data sets. The source code for this package has been uploaded onto GitHub (https://github.com/AlFontal/sdcpy), and a copy has been archived on the HELICAL community in Zenodo for long-term preservation (https://doi.org/10.5281/zenodo.4949813). Further expanding on accessibility, a web application has been developed to provide an interactive GUI for the package, thereby ensuring ease of use. The application can be found on GitHub (https://github.com/AlFontal/sdcpy-app), and is also readily available for use on the Heroku platform (https://sdcpy-app.herokuapp.com/).
- Association of ultrafine aerosols enriched in metals sourced from areas of intensive farming and urban pollution to Kawasaki Disease in Japan. Accepted Manuscript available at https://doi.org/10.1088/1748-9326/acd798. Accompanying code open

sourced at Github (https://github.com/AlFontal/kd-metals-swc) and deposited in Zenodo (https://doi.org/10.5281/zenodo.7948389)

- Assessment of the spatial coherence in the temporal dynamics of Kawasaki Disease at the prefectural level. Spatial clusters were observed not only during the early epidemic events (1979-1986), but also in current seasonal patterns. This provides a critical insight into the geographical and temporal factors influencing the disease's spread, furthering the evidence towards an environmental driver (manuscript in preparation).