Plan Overview

A Data Management Plan created using DMPonline

Title: The impact of climatic factors and climate extremes on pregnancy, child and adolescent health

Creator: Elena Raffetti

Principal Investigator: Elena Raffetti

Affiliation: Karolinska Institutet

Funder: Forte (Swedish Research Council for Health, Working Life and Welfare)

Template: Forte

ORCID iD: 0000-0001-8742-3986

Project abstract:

The study is based on the Developmental Origins of Health and Disease (DOHaD) study, a total population nationwide cohort including all individuals born in Sweden since 1973. The cohort comprises 5,751,581 births, with detailed information on perinatal outcomes and familial socioeconomic and health histories. This is made possible through record linkage of several national and regional health and administrative registers, including, for example, the Medical Birth Register, the National Patient Register, the Prescribed Drug Register, the Cause of Death Register, and the Longitudinal Integrated Database for Health Insurance and Labour Market Studies (LISA). Time series and case time series designs will be used to evaluate the effect of non-optimal temperature and precipitation pregnancy, neonatal and other health outcomes in children and adolescence (< 18 years of age). Using time series analysis, we will quantify the excess risk associated with non-optimal temperatures and precipitation. We will also explore the role of potential effect modifiers, such as maternal age, geographical location, and socio-economic determinants. Interactions with other climatic variables, exposure lags, and the fraction of events attributable to non-optimal temperature and precipitation.

ID: 176418

Start date: 01-05-2025

End date: 31-12-2027

Last modified: 25-05-2025

Grant number / URL: 2024-00833

Copyright information:

The impact of climatic factors and climate extremes on pregnancy, child and adolescent health

GENERAL INFORMATION

Project Title

The impact of climatic factors and climate extremes on pregnancy, child and adolescent health

Project Leader

Elena Raffetti

Registration number/corresponding

2023-01774

Version

V1

Date

01-01-2025 to 31-12-2028

DESCRIPTION OF DATA – REUSE OF EXISTING DATA AND/OR PRODUCTION OF NEW DATA

How will data be collected, created or reused?

The project will primarily reuse existing, high-quality register data from Swedish national and regional registers. These include the Medical Birth Register, National Patient Register, Prescribed Drug Register, Cause of Death Register, and LISA database, covering births in Sweden since 1973 (totaling over 5.7 million individuals). Climatic data (temperature, precipitation, and related variables) will be integrated from meteorological sources at a high spatial and temporal resolution to capture exposure profiles throughout pregnancy and childhood.
Data formats will vary, including tabular register data in standard statistical formats (e.g., CSV, SAS,

Stata) and gridded climatic data (netCDF, GeoTIFF). The volume of data is expected to be substantial due to the large cohort size and detailed temporal climatic data, but manageable within secure server infrastructure.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

All datasets will be accompanied by comprehensive metadata detailing data origin, structure, coding schemes, and collection methods, adhering to relevant metadata standards for epidemiological and environmental data. Register data are well-documented by the data providers, and climatic data will be annotated with source, temporal and spatial resolution, and quality indicators.

Data quality is ensured through the use of validated register sources and standardized data collection procedures. Quality assurance measures include cross-validation of exposures and outcomes, checks for missing or inconsistent data, and reproducibility of data extraction protocols.

DOCUMENTATION AND DATA QUALITY

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

The study data will be documented using comprehensive metadata records that describe the content, structure, and provenance of all datasets. For the register data, we will rely on existing detailed documentation provided by the Swedish national health and administrative registers, which include descriptions of variable definitions, coding schemes, update frequencies, and data collection methods.

For the climatic data, metadata will include information on data source (e.g., meteorological stations or modeled datasets), temporal and spatial resolution, units of measurement, data quality indicators, and processing steps applied (e.g., interpolation or aggregation).

All metadata will be formatted according to international standards such as the Data Documentation Initiative (DDI) for social science and epidemiological data, and ISO 19115 for geospatial data. This ensures that the data structure, variables, and collection methods are fully described for both current use and future reuse.

We will also maintain a data dictionary that defines all variables, coding values, and any derived variables created during data processing. Additionally, detailed protocols will be kept describing data linkage procedures, cleaning methods, and analytic dataset creation, ensuring transparency and reproducibility.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

Data quality will be ensured through multiple complementary approaches. The Swedish national registers used in this project are subject to rigorous quality control procedures at the point of data collection, including standardized reporting protocols and regular audits, which contribute to high completeness and validity of key variables such as birth outcomes and health diagnoses.

For climatic data, quality assurance includes the use of well-established meteorological data sources with documented calibration and validation processes. Where modeled or interpolated climate data are used, validation against observed measurements will be performed to assess accuracy.

Within the project, we will conduct consistency checks and validation steps during data processing, including verification of ranges, detection of outliers or missing values, and cross-referencing linked datasets to identify inconsistencies. Repeated measurements and longitudinal data points enable assessment of temporal trends and potential data anomalies.

All quality control procedures and results will be documented thoroughly in project logs and data processing protocols, ensuring transparency and reproducibility. Any data corrections or exclusions will be clearly recorded.

STORAGE AND BACKUP

How is storage and backup of data and metadata safeguarded during the research process?

All research data and associated metadata will be securely stored on KI’s dedicated research servers, which are protected by institutional IT security policies including firewalls, encryption, and multi-factor authentication and BIANCA server of UPPMAX. These servers comply with national regulations for handling sensitive personal data.

Regular automated backups of all data and metadata will be performed daily and stored on physically separate backup systems to protect against data loss due to hardware failure or other incidents. Backup integrity is routinely tested to ensure restorability.

Access to data storage systems will be restricted to authorized project members only, with role-based permissions controlling data read and write privileges. Data transfers within the project environment will use encrypted connections.

Throughout the research process, data versioning and clear documentation of changes will be maintained to safeguard data integrity and support reproducibility.

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

Data security is ensured through multiple layers of protection consistent with GDPR and Swedish data protection regulations. Sensitive and personal data are stored on secure servers with restricted physical and digital access controlled by [Your Institution]’s IT security policies.

Access to sensitive data is granted only to authorized project personnel based on the principle of least privilege, using authenticated accounts with strong passwords and, where possible, multi-factor authentication. Access permissions are regularly reviewed and updated.

Data are pseudonymized or anonymized whenever possible before analysis to minimize risks. Transfers of sensitive data within the research group are conducted via encrypted channels (e.g., SFTP, VPN).

All data handling and access are logged and monitored to detect any unauthorized activity. Staff receive mandatory training on data protection and confidentiality obligations.

LEGAL AND ETHICAL ASPECTS

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

Data handling in this project strictly adheres to all relevant legal frameworks, including the European Union’s General Data Protection Regulation (GDPR) and Swedish laws governing personal data and confidentiality. Prior to data access, necessary ethical approvals and data use agreements are obtained from relevant authorities and data custodians.

Personal data are handled with the highest confidentiality standards; sensitive information is pseudonymized or anonymized where possible, and data processing is limited to authorized personnel. The project implements robust technical and organizational measures to prevent unauthorized access, data breaches, or misuse.

Intellectual property rights related to the original register data remain with the data providers, and usage rights are governed by formal agreements that define permissible use, publication, and sharing conditions.

Throughout the project, data management complies with institutional policies and national regulations, and all staff involved are trained in legal and ethical requirements for data protection and confidentiality.

How is correct data handling according to ethical aspects safeguarded?

The project upholds the highest ethical standards in all aspects of data handling. Ethical approval has been obtained from the relevant regional ethical review board before accessing and processing any personal data.

Data are managed with strict respect for participant confidentiality and privacy, including the use of pseudonymization or anonymization to minimize re-identification risks. Data access is restricted to authorized personnel trained in ethical principles and data protection.

All research activities comply with the Declaration of Helsinki and applicable national ethical guidelines. Potential risks to individuals’ privacy and well-being are continuously assessed and mitigated throughout the project lifecycle.

Transparency and accountability are ensured by maintaining detailed documentation of data handling procedures and decisions, enabling audit and review if required.

ACCESSIBILITY AND LONG-TERM STORAGE

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

Research data and associated metadata will be made accessible in accordance with the FAIR principles, ensuring that data are Findable, Accessible, Interoperable, and Reusable. Metadata describing the datasets will be published in a secure, controlled-access repository hosted by [Your Institution] or a national data archive such as the Swedish National Data Service (SND).

Due to the sensitive nature of personal and health-related data, direct access to raw individual-level data will be restricted and only granted to authorized researchers who comply with data use agreements and ethical approvals. Access requests will be subject to review by the data governance body to ensure compliance with legal and ethical requirements.

An embargo period may be applied during the initial analysis phase to allow for primary publication of results by the project team. After this period, anonymized and aggregated datasets, along with detailed metadata, will be made available for secondary research purposes under controlled conditions.

All data sharing will comply with GDPR and other relevant regulations, and users must acknowledge the data sources and respect intellectual property rights as specified in the data use agreements.

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

Long-term storage of research data and metadata will be secured through collaboration with [Your Institution]'s research data management services and/or a recognized national repository such as the Swedish National Data Service (SND). These entities provide reliable, sustainable digital preservation infrastructures that ensure data integrity, security, and accessibility over time.

Data selected for long-term preservation will primarily include anonymized, aggregated datasets, detailed metadata, and documentation sufficient to enable future reuse and replication of the research. Selection criteria will focus on data that have lasting scientific value, comply with legal and ethical requirements for storage, and are feasible to preserve without compromising participant confidentiality.

The responsibility for overseeing long-term storage will transition from the project leader to the institutional or national data repository, which follows established best practices for data curation and preservation.

Regular reviews will be conducted to ensure stored data remain accessible and usable with evolving technologies, including plans for format migration and documentation updates as needed.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

Yes, specific software and computational environments will be necessary to access, understand, and analyze the research data over the long term. The datasets, which include detailed epidemiological registers and geospatial climatic data, require statistical software such as R, Stata, or SAS for analysis, alongside GIS tools (e.g., QGIS or ArcGIS) for spatial data handling.

The project utilizes the **BIANCA UPPMAX server**—a high-performance computing resource provided by Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX)—to perform large-scale data processing and analyses. To ensure reproducibility and future usability, all analysis scripts, source code, and software versions used on BIANCA UPPMAX will be archived with thorough documentation.

Where feasible, open-source software and standard data formats (CSV, NetCDF, GeoTIFF) will be preferred to maximize accessibility. Additionally, containerized environments (e.g., Docker) or virtual machine images may be created to preserve computational workflows and software dependencies, supporting reproducibility regardless of future changes in software platforms.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

To ensure the findability and citability of research data, unique and persistent identifiers such as Digital Object Identifiers (DOIs) will be assigned to datasets and key metadata records. This will be done through collaboration with recognized data repositories such as the Swedish National Data Service (SND) or institutional repositories that support DOI registration.

The assignment of DOIs facilitates unambiguous referencing, tracking of data usage, and promotes compliance with FAIR data principles. Metadata linked to these DOIs will be maintained and updated as needed to reflect any changes in data availability or access conditions.

Long-term management and resolution of these identifiers will be safeguarded by the repository organizations, ensuring stable and persistent access to the data over time.

RESPONSIBILITY AND RESOURCES

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

During the research project, the Principal Investigator (PI) holds overall responsibility for data management, ensuring compliance with legal, ethical, and institutional requirements. The PI is supported by Elena Raffetti, who provides expert guidance and coordination on data handling, quality assurance, and documentation.

Shivang Pandey, the PhD student, is responsible for the day-to-day management of the data, including collection, processing, documentation, and ensuring adherence to the data management plan throughout the project lifecycle.

After the project concludes, long-term data management and storage responsibilities will be transitioned to KI's dedicated research data management team and/or a recognized national repository such as the Swedish National Data Service (SND). The PI will oversee that proper handover procedures and documentation are in place to support sustainable data preservation and future accessibility.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

The data management activities for this project will require dedicated resources in several areas:

- **Storage and Backup:** Costs related to secure storage on institutional servers and backup infrastructure, including daily automated backups and offsite replication, will be covered by [Your Institution]'s research IT budget or through national infrastructure grants (e.g., UPPMAX). The volume of data, including large register and climate datasets, necessitates scalable storage solutions.
- **Labour Input:** Considerable personnel time will be invested by the PhD student (Shivang Pandey) and supporting staff for data cleaning, processing, documentation, and metadata creation. The PI and data management expert Elena Raffetti will allocate time for oversight, quality assurance, and coordination with data repositories and ethical/legal compliance bodies.
- **Access Provision and Security:** Resources will be allocated for managing controlled access systems, user authentication, and monitoring, ensuring compliance with data protection regulations.

- **Long-Term Storage and Preservation:** Costs may include fees associated with depositing data and metadata in national repositories like the Swedish National Data Service (SND) or institutional archives, as well as periodic data format migrations and system maintenance to safeguard future usability.

To ensure that data fulfil the FAIR principles, resources will be dedicated to:

- Creating rich, standardized metadata following international standards.
- Using interoperable and open data formats wherever possible.
- Implementing persistent identifiers (DOIs) for datasets and metadata.
- Documenting data provenance, processing workflows, and software environments to facilitate reuse.
- Training project staff on FAIR data management practices and maintaining up-to-date documentation and data dictionaries.