Plan Overview

A Data Management Plan created using DMPonline

Title: ENVIRONMENTAL AND OCCUPATIONAL NOISE EXPOSURE IN RELATION TO INCIDENCE OF TYPE 2 DIABETES, INCLUDING THE ROLE OF AIR POLLUTION AND METABOLIC SYNDROME COMPONENTS

Creator:Göran Pershagen

Principal Investigator: Göran Pershagen

Data Manager: Jessica Edstorp

Affiliation: Karolinska Institutet

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0000-0002-9701-1130

Project abstract:

Large parts of the population are exposed to traffic noise, particularly in urban areas, and high noise levels occur in many workplaces. The aims of this project are to estimate exposure-response relationships for incidence of type 2 diabetes (T2D) related to long-term exposure to road traffic, railway and aircraft noise, as well as occupational noise. Furthermore, mediation by sleep disturbances and overweight/obesity is investigated, as well as interactions with air pollution and clinical biomarkers, to elucidate important etiological pathways. The project is based on pooled analyses of nine Scandinavian cohorts, totally including more than 300 000 individuals. Detailed longitudinal exposure to traffic noise from different sources, air pollutants and greenness, as well as occupational exposures, are estimated with state-of-the-art methods and supplemented with questionnaire and registry data on risk factors as well as clinical measurements. Incidence of T2D is assessed by combining information from medical examinations, high quality registers, questionnaires and biomarker measurements. Population attributable risks are estimated by combining population data on exposure with information on exposure-response relationships. In several aspects the project is unique and addresses questions which have never been studied before. Our results will provide important guidance for prioritization of preventive measures to promote health sustainable urban development and safe workplaces.

ID: 152870

Start date: 01-06-2024

End date: 31-12-2026

Last modified: 29-05-2024

Grant number / URL: 2023-02077

Copyright information:

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

ENVIRONMENTAL AND OCCUPATIONAL NOISE EXPOSURE IN RELATION TO INCIDENCE OF TYPE 2 DIABETES, INCLUDING THE ROLE OF AIR POLLUTION AND METABOLIC SYNDROME COMPONENTS

General Information

Project Title

ENVIRONMENTAL AND OCCUPATIONAL NOISE EXPOSURE IN RELATION TO INCIDENCE OF TYPE 2 DIABETES, INCLUDING THE ROLE OF AIR POLLUTION AND METABOLIC SYNDROME COMPONENTS

Project Leader

Göran Pershagen

Registration number/corresponding

2023-02077

Version

Version 1

Date

2024-05-29

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

Study subjects
Nine Scandinavian cohorts constitute the sampling frame for the study. Four cohorts were recruited in Stockholm County: [Stockholm Diabetes Preventive Program (SDPP), 60-year old men and women (Sixty), Twin Registry (SALT) and Swedish National Study of Aging and Care in Kungsholmen (SNAC-K)], with a total of more than 22 500 participants. The Malmö Diet and Cancer (MDC) study recruited 28 098 men and women living in the city of Malmö. The Danish Diet Cancer and Health (DCH) cohort enrolled 57 053 subjects from the greater Copenhagen or Aarhus areas. Finally, the Danish Nurse Cohort (DNC) included 28 731 female nurses from the whole of Denmark. Overall, enrollment focused on ages 35 to 99 years and occurred 1992-2004. The study subjects of these seven cohorts formed the NordSOUND project, and answered questionnaires at recruitment on lifestyle factors, health status and socioeconomic characteristics. Blood samples were obtained from participants in the Swedish cohorts. Anthropometric measurements were performed by trained nurses for all cohort participants, except SALT and DNC, which used self-reported data. In addition, two cohorts are included: The Swedish Mammography Cohort (SMC) with 13 680 women residing in the Uppsala area, who completed a self-administrated questionnaire concerning diet and alcohol intake as well as on risk factors for breast cancer. Questions on anthropometric markers were also included. Subsequent questionnaires every tenth year provided more detailed information on lifestyle factors as well as on medication use, sleep habits, family disease history, stress and social support. The Danish National Health Survey (DNHS) cohort included 177 639 subjects randomly selected across Denmark who answered a questionnaire on cohabiting status, education, occupation, smoking, alcohol consumption, diet, physical activity and anthropometric data. Information on age, sex, education and labor market affiliation was obtained from national registers.

Exposure assessment

Traffic noise: Transportation noise exposure is assessed based on well validated models. Road traffic and railway noise are modelled using the Nordic Prediction Method or an update of this method. For road traffic noise the input variables include geocodes, screening by terrain and buildings, and information on annual average daily traffic, distribution of light/heavy traffic, travel speed, and road type for all major road links. Railway noise is calculated for all addresses within a 1000 m buffer around all railway tracks. Input variables include geocodes, screening by terrain and buildings, and average number of trains per period (day/evening/night), train types, and travel speed. In addition, cities with trams and/or metro include these in the calculations. Aircraft noise is estimated using noise maps obtained from local or national authorities. Noise exposure from airports and airfields is modelled using the Danish Airport Noise Simulation Model or the Integrated Noise Model 7.0. Detailed noise assessments have been performed every fifth or tenth year and noise levels for the years between those with estimates are calculated based on linear interpolation or other approximation methods. For each participant, the time-weighted average noise exposure from each traffic noise source during follow-up is calculated, taking into account all addresses where the subject has lived, and considering the duration of residence at each address. In addition, combined exposure to multiple traffic noise sources is estimated.

Air pollution: Levels of air pollution are estimated at all residential addresses during the study period for the subjects in the nine cohorts, using validated high-resolution dispersion models. Air pollution exposure is represented by PM2.5, which is influenced by both local and long-range transport, and by NO2, primarily reflecting local emissions, such as from road traffic. Interpolation of air pollution levels between years with assessments as well as calculation of individual time-weighted exposures are done using similar methodology as for transportation noise.

Occupational exposures: This is focused on noise and combustion particles, based on occupations of the study subjects combined with information from a job-exposure-matrix (JEM). Occupational noise exposure is estimated based on a JEM developed in Sweden. The JEM is based on occupational measurements and specifies the annual average of the daily 8-hour equivalent A-weighted sound pressure level in five exposure classes. The noise level is matched on time period since noise levels differ within an occupation across time. Occupational noise exposure at recruitment is used and, if available, exposure in certain time-windows during follow-up. Occupational exposure to combustion particles is handled in the same way as noise, but based on an adapted Finnish JEM.

Covariates: Selection of covariates is done a priori, based on existing literature, biological plausibility, and availability of harmonizable variables across cohorts. Cohort participants filled in questionnaires at recruitment with dietary and lifestyle variables, including smoking status, smoking intensity, alcohol consumption and leisure-time physical activity. The questionnaires also provided information on sleep disturbances. Individual educational level and marital status are obtained from national registers or questionnaires, and area-level (small socioeconomically homogeneous areas with around 1000-2000 inhabitants) mean income from registers. Green areas are assessed from satellite images, primarily using the normalized difference vegetation index.

Outcome assessment

Incidence of type 2 diabetes (T2D): All relevant information collected within each cohort is used to identify prevalent cases of diabetes at baseline, who will be excluded from the longitudinal analyses, and incident cases during follow-up until 2020. This includes linking with the Patient and Prescribed Drug Registers as well as using self-reported diabetes in the questionnaires and biomarker measurements (primarily fasting glucose and HbA1c), both at baseline and during follow-up. The methodology for identification of T2D cases has already been used successfully for cohorts in our project. In Sweden the Patient Register has full coverage since 1987 but contains comprehensive outpatient data only since 2001 and the Prescribed Drug Register was started in 2005. In Denmark both the Patient and Prescribed Drug Registers have full coverage during virtually the whole follow-up period of their cohorts. In addition, there are national T2D registers in Sweden and Denmark, but they do not have comprehensive coverage during most of the follow-up period. Overall, some registry sources for identification of T2D cases are lacking, primarily for Swedish cohorts during the early part of the follow-up period. However, to the extent that this is unrelated to the exposures under study it will not affect the validity of the findings.

Anthropometry: In a majority of the cohorts, measurements were performed by trained nurses of height, weight and waist circumference at recruitment, while corresponding information was self-reported in the remaining cohorts. In two cohorts repeated measurements were performed during the during the follow-up period, enabling longitudinal assessment of anthropometric characteristics. In the mediation analyses, overweight/obesity data based on anthropometric information at recruitment will be combined with incident T2D during follow-up.

Clinical biomarkers and measurements: For five of the Swedish cohorts information at recruitment of study subjects is available on blood pressure based on measurements by trained nurses, as well as on clinical biomarkers, including serum glucose and lipids. In the SNAC-K cohort measurements of glycated hemoglobin (HbA1c) levels were performed every 3-6 years from 2001 to 2019. For SDPP participants oral glucose tolerance tests have been made at three different occasions during follow-up. The biomarker information enables accurate determination of T2D and prediabetes, with due consideration of treatment, and will also be used for validation of the registry and questionnaire-based information, primarily to determine the degree of underdiagnosis.

Population attributable risks: Risk assessment is based on estimates of the population exposure to transportation noise and occupational noise in the catchment areas of the participating cohorts (Aarhus, Copenhagen, Denmark, Malmö, Stockholm and Uppsala) using the high-resolution modeling techniques described above. This will be combined with exposure-response functions obtained in the project to estimate the number of cases of T2D attributable to transportation and occupational noise. In particular, assessment is made of the number of cases related to interactions between traffic noise and air pollution as well as with occupational exposures.

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

Information on T2D from registers and individual lifestyle characteristics from questionnaires as well as results from anthropometric measurements is primarily stored at the universitiy departments where the participating cohorts in this project originate. For Karolinska Institutet this includes the Institute of Environmental Medicine (SIXTY, contact person: Karin Leander) and departments of Global Public Health (SDPP, contact person: Hrafnhildur Gudlonsdottir), Medical Epidemiology and Biostatistics (SALT, contact person Patrik Magnusson) and Neurobiology, Care Sciences and Society (SNAC-K, contact person Debora Rizzuto). Corresponding information for SMC and the Danish cohorts is stored at Uppsala University, Copenhagen University and the Danish Cancer Registry.

Biological samples are stored in biobanks at Karolinska Institutet/Region Stockholm and Lund University/Region Skåne, Halland, Blekinge, Kronoberg). No biological data are used for SMC or the Danish cohorts. Environmental data for the Stockholm and Uppsala cohorts are added by the Institute of Environmental Medicine based on residential history, while corresponding information for the Danish cohorts is generated at the departments mentioned above. Data enabling personal identification of study subjects are primarily stored at the departments where the cohorts are administered.

Data with CPR-number from the Danish cohorts and personal ID-number for the Swedish cohorts are uploaded through a secure server at Statistics Denmark ("The Database"). The Database pseudonymizes the data using a key variable only known by the Database. The pseudonymised data are placed in a specific project folder, were only approved users have access. Consequently, the user will never get direct access to the CPR- or personal ID-numbers.

The data will be primarily stored in Excel, CSV and STATA files, with an estimated maximum total size of 25 GB. All data will be accompanied by comprehensive metadata descriptions to ensure clarity and ease of use. These descriptions will include:

- Data structure: Detailed variable descriptions, data types, and relationships between datasets.
- Collection methods: Descriptions of data collection methodologies, including sources, tools used, and timeframes.
- Processing and transformation: Records of any data cleaning, processing steps, and transformations applied.

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.?

All data used in the project will be documented with detailed metadata, encompassing variable descriptions, data formats, and the context of data collection. The documentation will include:

• Study protocols: Detailed outlines of the study design and methodologies.

• Codebooks: Comprehensive descriptions of all variables and their values.

• Logbooks: Records of all data handling activities, ensuring transparency and reproducibility.

• Program files and scripts: Documentation of all code used for data processing and analysis, stored in accessible formats such as Do-files for STATA or r-files for R.

• Output files: Results from data analyses, with accompanying descriptions to facilitate interpretation.

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

Data quality will be safeguarded through rigorous documentation and validation procedures, including:

• Quality Assurance: All data handling and analysis activities will be documented using the KI ELN (Electronic Lab Notebook) or equivalent systems that meet KI's information safety standards.

• Data Validation: Comprehensive validation checks will be performed to verify the accuracy of data input and processing steps.

Storage and backup

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

All data utilized in the project will be stored on the secure central file server of Statistics Denmark (DST). The platform ensures:

• Regular Backups: Automated regular backups to prevent data loss.

• Data Security: Compliance with high standards for data security, including encryption and secure access protocols.

• Controlled Access: Access to data is strictly regulated, with permissions granted only to authorized personnel.

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 and controlled access will be ensured through several measures:

• Strict access control: Only authorized personnel will have access to the data. Permissions are managed to ensure that sensitive data are only accessible to those who need it.

• Pseudonymization: Personal identifiers (person-keys) will be pseudonymized and stored separately from the personal data to

enhance privacy.
• Data encryption: Sensitive data will be encrypted to prevent unauthorized access.
• Monitoring and logging: All data access and handling activities will be logged and monitored to promptly detect and address any security breaches.

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 will comply with all relevant legal requirements, including:
• Regulations compliance: All project staff are trained and informed about the legal requirements and institutional policies at KI and other participating institutions regarding personal data handling.
• Confidentiality agreements: All personnel will sign necessary agreements according to policies of the participating institutions and DST.dk to ensure the protection of personal data.
• Intellectual property rights: Data usage will hold to the intellectual property policies of the participating institutions and the terms outlined in the project's data management plan.

How is correct data handling according to ethical aspects safeguarded?

The project has received all required ethical permissions. All scientists and staff involved in the project are informed about the conditions for data handling according to the ethical permissions.

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?

A Disclosure Agreement (the "Agreement") has been signed between the Institute of Environmental Medicine, Karolinska Institutet ("Provider") and the Danish Cancer Society ("Recipient"), collectively referred to as the "Parties" and individually as a "Party", regarding data accessibility and storage:
The Parties are obligated to assess the risks to the rights and freedoms of natural persons posed by each Party's processing and to implement measures to address these risks. Depending on their relevance, this may include:
a) pseudonymization and encryption of personal data;
b) the ability to ensure the continued confidentiality, integrity, availability and resilience of processing systems and services;
c) the ability to restore the availability of and access to personal data in a timely manner in the event of a physical or technical incident;
d) a procedure for regular testing, assessment and evaluation of the effectiveness of the technical and organizational measures to ensure secure processing.
Each Party must implement appropriate technical and organizational measures, taking into account the current state of the art, implementation costs and the nature, scope, context and purposes of the processing, as well as the risks of varying likelihood and severity to the rights and freedoms of natural persons, in order to ensure a level of security appropriate to those risks and in accordance with Article 32 of the General Data Protection Regulation.
Provider guarantees that Provider has the necessary legal basis to disclose the data to Recipient. Recipient guarantees that
a) Recipient has the necessary legal basis to collect data from Provider and process the collected data;
b) data will be processed solely for statistical or scientific purposes and will not be disclosed for processing for any other purpose;
c) the data is necessary for statistical or scientific processing. Data which, after collection, prove to be unnecessary for statistical or scientific processing must be deleted, destroyed or returned as soon as possible;
d) disclosure by Provider is made only to persons of Recipient who are authorised by Recipient to access the data concerned. Recipient has only authorised persons handling the personal data. Individuals are not authorized for uses for which they have no need;
e) necessary instructions are given to the employees of Recipient who have access to the personal data. In this context, the employees are informed that the personal data may solely be processed for statistical or scientific purposes and that the personal data may not later be processed for purposes other than scientific or statistical ones;
f) the dissemination of the results of Recipient's data processing is done in such a way that it is not possible to identify individuals;
g) at the end of the examination, or when it is not relevant to examine further, the data will either (a) be deleted, anonymised,

destroyed or returned in such a way that it subsequently is not possible to identify natural persons from the data or in combination with other data, or (b) be transferred for archiving in accordance with the legislation on archives;
h) appropriate security measures are implemented. The guarantee is given from the time of the transfer and is maintained until termination of the Agreement.
As indicated above, all access to data in sever files is strictly regulated. Only authorized persons have access to data. Sensitive personal data such as person-key are only available to 2-3 database managers. Access to data is determined by PIs for the different cohorts involved in the project together with the project PI. This also relates to future use of the data generated within the project.
Some metadata will be available and can be shared, also post publication. However, unique data possibly revealing identity, such as address information or geocodes, will not be available. This also relates to individual health data.

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 will be managed as follows:
• Decision making: The selection of data for long-term storage will be determined by the Principal Investigators (PIs) of the involved cohorts, in consultation with the project PI.
• Storage location: Data will be securely stored on the central file server at Statistics Denmark.
• Future use: Decisions regarding future use and accessibility of the stored data will be made by the PIs, ensuring compliance with all ethical and legal requirements.

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?

For long-term data usability, the following will be ensured:
• Analysis documentation: All analysis information will be stored as (syntax code, Stata or R) in accessible text formats to facilitate future use and replication of analyses.
• System compatibility: Documentation will be maintained using the KI ELN or comparable systems that ensure long-term accessibility and compliance with data standards.

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

DOIs will be generated and maintained in collaboration with KI library, also when data sets are stored in data repositories.

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?

Göran Pershagen (goran.pershagen@ki.se) is responsible for the overall data management structure, while Jessica Edstorp (jessica.edstorp@ki.se) handles the day-to-day data mangement activities. Since Göran Pershagen has retired, but continue to work (50%) as professor at KI during the project period, the group leader Petter Ljungman (petter.ljungman@ki.se) will be responsible for long-term storage of data from the project.

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 Swedish Research Council granted only 40% of the total costs applied for in the project. Consequently, a discussion is ongoing on prioritization of the project activities. However, ample resources will be secured for data management and storage.