Plan Overview

A Data Management Plan created using DMPonline

Title: The effect of obesity on cardiovascular disease: mediation analysis with Mendelian randomization

Creator: Arvid Sjölander

Principal Investigator: Arvid Sjölander

Data Manager: Arvid Sjölander

Affiliation: Karolinska Institutet

Template: Template for Swedish Research Council

ORCID iD: 0000-0001-5226-6685

Project abstract:

Obesity refers to an excessive accumulation of body fat, and is often defined as a body mass index (BMI; weight[kg]/height[m]2) above 30. It is one of the most common health-threatening conditions in the world, with a global prevalence of 13%. A particularly negative consequence of obesity is the highly elevated risk of cardiovascular disease (CVD), which, according to the World Health Organization (WHO), is the leading cause of death world-wide [1]. The effect of obesity on CVD is potentially mediated through a wide range of biological markers and conditions (e.g. high blood pressure, high serum cholesterol and glucose intolerance), but the relative importance of these mediators is not well understood. In the last two decades, many statistical methods have been developed for mediation analysis [2]. However, most of these methods crucially rely on the assumption of no unmeasured confounding, i.e. that the exposure and mediators can be considered randomized within levels of observed covariates. This assumption is often highly dubious in studies of obesity and CVD, since these conditions are influenced by a large number of genetic and environmental factors, of which many are typically unmeasured. A popular way to deal with unmeasured confounding is to use Mendelian randomization [3]. This method utilizes a genetic variant (or a set of variants), which is known to influence the exposure of interest, and which can be considered "randomized by nature", to infer the causal exposure effect on the outcome. Recently, Mendelian randomization methods have been used successfully in several studies, to estimate the effect of obesity on CVD. However, limited Mendelian randomization methods exist for mediation analysis. The specific aims of this cross-disciplinary project are to 1. develop Mendelian randomization methods for hypothesis testing in mediation analysis. 2. develop Mendelian randomization methods for effect estimation in mediation analysis. 3. implement the proposed methods into a professional software package, and make it publicly available to other researchers over the world. 4. use the developed methods to explore the mediating pathways from obesity to CVD. For this aim we will use data from the UK biobank, which is one of the largest biobanks in the world.

ID: 67341

Last modified: 11-01-2021

Grant number / URL: 2020-01188

Copyright information:

The effect of obesity on cardiovascular disease: mediation analysis with Mendelian randomization

Description of data

How will data be collected, created or reused?

We will not collect any new data but use a large prospective cohort that is already collected; the UK biobank. The UK Biobank is a prospective cohort collected in 2006-2010 from all over UK (https://www.ukbiobank.ac.uk/). 500,000 people aged 40-69 years were subjected to a health check where blood was drawn for biomarkers, questionnaires collected, including information on medical treatments. Data have been linked to mortality-, and cause-specific registers as well as national hospital electronic data. The UK Biobank has been genotyped by an Affymetrix array with genome-wide coverage and full imputation is available.

What types of data will be created and/or collected, in terms of data format? Include version numbers if applicable.

We will not collect any new data, see information above, nor create any new data. New data will be generated in form of Codebooks, Logbooks, Analysis plans Syntax scripts and Output files from database systems and statistical software.

What volumes of data will be created and/or collected?

- < 100 GB

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, file naming-format-versioning, etc

Documentation of the material follows the approved guidelines of the Department of Medical Epidemiology and Biostatistics. These include a standardized folder structure for documentation comprising of Codebooks (metadata about data), Logbooks (metadata about data processing and cleaning), Analysis plans (including detailed descriptions of the data retrieval and research studies), Manuscripts, syntax scripts and output files from database systems and statistical software (for data management and analysis), Program folders, Data folders and Communications with data providers. The department also has a standard for variable naming and coding for primary data collections.

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

Quality documentation will be included by the data provider. Quality controls will be performed at delivery of the data to ensure that the delivered data is correct.

Storage and backup

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

- KI Server

Access to storage of data is guarded strictly by IT-policy at the department with different levels of authorization given to a user (researcher/nonresearcher) on the PI’s approval. The department’s research data and other storage are backed up every day with snapshots of different versions available to recall. For large-scale genotype data in UK biobank we use the Swedish National Infrastructure for Computing (SNIC) at Uppmax in Uppsala where data are stored, with backup and safety protocols applied, see information below. The genotype data from UK biobank are only located at SNIC Uppmax servers, and never transferred to MEB. The files are too big and cannot be analyzed without a computer cluster. The phenotypes are thus saved at MEB´s server Vector.

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

For this specific project, all archived data (i.e. from surveys, in-person testing, genotypes from archived DNA and blood samples) from the cohort will be held strictly confidential with pseudonymized data. Sensitive large-scale data, such as genotypes, are kept on a high-security computer cluster system provided by SNIC. This resource is developed specifically for sensitive data in accordance with the Swedish Personal Data Act, and supported by the Swedish Research Council to provide bioinformatics support to life science researchers in Sweden. The basic layout of the system is a secure centralised service at SNIC-UPPMAX where data are stored. Within the secure environment a High Performance Computing (HPC) resource is provided for large jobs. All projects on the system will have their own set of virtual machines. Users access their virtual machines via a secured graphical terminal using two-factor authentication, and all communication is encrypted. This way the data are controlled and secured in the central secure vault and the risk of losing data is minimized. The environment is secured by a strict firewall and no internet access is allowed within the environment. Import and export is handled by a secured file transfer service. Only project Principal Investigators are allowed to export data from the secured system. Also, as stated below, only a local database manager responsible for keeping track of data collections and linkages to registers have access to the personal identity numbers of the study participants.

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?

KI as an organization complies with GDPR in both legal and ethical aspects. More information can be found at www.h ttps://medarbetare.ki.se/gdpr
All identifying information such as name and personal identity number have been removed before data sets are made available to personnel for analysis; we only work with pseudonymized data. Only a local database manager responsible for keeping track of data collections and linkages to registers have access to the personal identity numbers of the study participants. We have applied to use the UK Biobank Resource under Application Number 22224.

How is correct data handling according to ethical aspects safeguarded?

The UK biobank cohort has an ethical approval in the UK. No ethical approval is needed in Sweden to study the data from the UK biobank cohort.

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, licenses and limitations on the access to and reuse of data?

The data and all material is archived by the IT section at the department as per the archiving guidelines at the Department of Epidemiology and Biostatistics, and is made accessible whenever required (legally and ethically). No one is given access to the archived material without legal and ethical permissions, which are in general sought through the university’s registrar office. Also, since the data is subject to GDPR, only metadata is published openly, underlying data is made available upon request after ensuring compliance with relevant legislation and KI guidelines.

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

As described above the data are archived by the IT at the department and safeguarded with no access given to any user unless permitted legally and ethically by the registrar’s office at the university. The archiving guidelines include instructions for selection of files necessary to ensure reproducibility of published results, as well as safeguarding the use and readability of valuable data for future research. This includes ensuring that data, metadata and other documentation are saved in stable data file formats over time. For genotype data stored by SNIC, they have safeguarded storage of large-scale data and terms of use for academic reasearch.

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

Materials that are used for the data management, analysis and results are stored in a readable format,

in order to understand, partake of or use/analyse data in the long term, as according to departmental documentation guidelines. In addition, SNIC is required, which is alluded to above.

How will unique and persistent identifiers for the research data, such as a Digital Object Identifier (DOI), be obtained?

The university has a central database with DOIs of all the published articles, which is backed up regularly.

Responsibility and resources

Who is responsible for data management while the research project is in progress?

For this project, the principal investigator is responsible for the research project including data management when the project is ongoing. For the UK biobank data, database managers in the UK provide the data that we analyze in Sweden, at the department.

Who is responsible for data management, long-term storage after the research project has ended?

For all projects run at the department, code and data are stored according to our data documentation standards. Hence, all projects are reproducable based on the stored information even after the project is finished. For the genotype data stored at SNIC, they have local database managers and an excellent IT support system to maintain long-term support.

What resources (costs, labour or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)?

The important resources needed is IT infrastructure with long-term support, such as SNIC.

What resources will be needed to ensure that data fulfil the FAIR principles?

No particular additional resources will be required will be needed to ensure that the data fulfil the FAIR principles.