Plan Overview

A Data Management Plan created using DMPonline

Title: EDCEL: Exploiting new Data to examine Celiac disease comorbidity

Creator: Jonas Ludvigsson

Principal Investigator: Jonas Ludvigsson

Data Manager: Jonas Ludvigsson

Affiliation: Karolinska Institutet

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0000-0003-1024-5602

Project abstract:

Celiac disease (CD) is a lifelong immune-mediated disorder that is characterized by villus atrophy and small intestinal inflammation. The disease occurs in genetically predisposed individuals exposed to oral gluten, and has been linked to both mortality and other complications, and is associated with a number of disorders. During the 1990s and early 2000, CD increased and has now established itself as one of the most prevalent chronic gastrointestinal disorders, occurring in 1 in 44 females and 1 in 72 males across a lifetime. Recent data suggest CD is an economic burden in many patients, but will also result in substantial workloss in affected patients. Earlier studies on the long-term prognosis of CD have often been limited to the use of hospital-based outcome measures (typically retrieved from the Swedish Patient register). In this proposal we will use new data sources to better understand the complications of CD: - a) Data on biopsy-verified CD from the ESPRESSO cohort will be linked to the SCREAM cohort covering the full population of the Stockholm region – roughly corresponding to 25% of the Swedish population (i.e. 2.5 million people). Through the SCREAM project we will access laboratory data that will serve as markers of early liver-kidney disease. - b) We will take advantage of newly collected primary health care data from 20 out of 21 Swedish regions. This will allow us to examine psychiatric/neuro-cognitive disorders typically cared for in primary care in addition to those comorbidities that are based on in- and outpatient hospital data.

ID: 71185

Start date: 01-01-2024

End date: 31-12-2026

Last modified: 23-05-2024

Grant number / URL: 2023-02207

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

EDCEL: Exploiting new Data to examine Celiac disease comorbidity

General Information

Project Title

EDCEL: Exploiting new Data to examine Celiac disease comorbidity

Project Leader

Jonas F Ludvigsson

Registration number at the Swedish Research Council

2023-02207

Version

1

Date

15 Jan, 2024

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

How will data be collected, created or reused?

Routine data from histopathology reports from Swedish pathology departments have been collected. Data are digital. They will be re-used. Data concern the gastrointestinal tract.

Individuals with a gastrointestinal biopsy report have been linked to the national Swedish Health register. Through the Statistics Sweden, matched controls, and first-degree relatives have also been identified and matched to health registers.

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

This is described in the study review: https://pubmed.ncbi.nlm.nih.gov/30679926/

Data from the 28 pathology departments will be stored in Excel and SAS format.

Data from national registers from Socialstyrelsen and SCB will be stored in SAS and SQL format.

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.?

A documentation of the data was produced in April 2019, and contains the following headlines. This documentation is continuously updated.

Contents

Brief information about the Histoepidemiology project (ESPRESSO).
Background
Aim
Brief information about the data (ESPRESSO)
Other datasets that were linked to ESPRESSO
Quality insurance of ESPRESSO
Data sources
More about Data clearing in ESPRESSO using information from SCB
Folder structure
WORK DESCRIPTION for Histoepidemiology study
Access to ESPRESSO in SQL
Persons in charge
Statisticians
Database administrator tasks
Datasets in SQL
ESPRESSO Codebooks

A codebook has been constructed by database manager Mariam Lashkariani.

All studies are preceded by analyses plans.

Working files are clearly labelled using dates (rather than version suffices)

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

Quality controls were applied when we received data from Swedish Pathology departments and from the National Healthcare registers, and will be applied again when the data are updated.

Quality controls included checking dates for emigration/death and birth in relation to biopsy dates. Quality controls also included checks of sex proportion, age distribution, but we also plan to re-run some important studies in the new dataset to see that results align with similar Swedish data.

The completeness of the linked data is assured by relevant government committees (Socialstyrelsen and SCB).

Storage and backup

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

Access to and storage of data is guarded strictly by IT-policy at the department with different levels of authorization given to a user (researcher/non-researcher) on the PI’s approval.

Data will be stored in the P:\ at the central server of MEB (department).

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

The department’s research data and other storage are backed up every day with snapshots of different versions available to recall.

Also access to the data saved on the server is restricted to group members/authorized personnel.

Data cannot be shared with outside researchers due to limitations of the ethics approval and the approvals of the different pathology departments, but the same data can be retrieved from the national healthcare registers and through contacting the participant pathology departments.

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. Handling of data is always done decoupled from identifiers. No identifiable data is published or shared publicly. Data is kept at secure servers within the KI firewall.

The key that can be used to identify individuals is only available to the relevant Swedish Government agency handling the matching.

More information can be found at www. https://medarbetare.ki.se/gdpr

How is correct data handling according to ethical aspects safeguarded?

Handling of data is always done decoupled from identifiers. No identifiable data is published or shared publicly. Documentation of ethical approvals (applications, amendments and decisions) and informed consent forms for the project are stored in the project folder electronically (and in the paper archive, if relevant). Ethical approvals are registered in the diary.

Linked data have been pseudonymized. The code key for pseudonymized data is kept by Statistics Sweden ([https://www.scb.se/)](https://www.scb.se/)

No physical data are stored. (no paper CRFs are stored)

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?

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 & ethical permissions, which are in general

sought through the university’s registrar office.
Because of the presence of personal data, only metadata will be published and shared openly

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 is archived by the IT at the department and safeguarded with no access given to any user unless permitted legally and ethically.
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.
The data will be stored until the PI Ludvigsson retires.

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?

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.
The following software are used to analyse data: SAS; Stata; R; and SPSS

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

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 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?

Ludvigsson (PI) has the overall responsibility for data management and long-term preservation of data while the research project is in progress.
Ludvigsson also has the overarching responsibility for the above when the research project has ended.

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?

We are continuously up-dating the storage and back up facilities at MEB. This is done centrally through core funding and through funding by Ludvigsson.
We expect the DMP will help other researchers identify our data source. We have also described our database through a review paper (cited earlier in the DMP), and in each paper based on this source we mention its

name "ESPRESSO".
We have kept the variable names used by Socialstyrelsen and SCB to make the dataset interoperable, and easy to understand.

Planned Research Outputs

Interactive resource - "Older Age of Celiac Disease Diagnosis and Risk of Autoimmune Disease: A Nationwide Matched Case-Control Study"

Scientific paper: Conditionally Accepted. J Autoimmunity 2024.

Planned research output details

Title	DOI	Type	Release date	Access level	Repository(ies)	File size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Older Age of Celiac Disease Diagnosis and Risk of ...		Interactive resource	Unspecified	Open	None specified		None specified	None specified	No	No