Plan Overview

A Data Management Plan created using DMPonline

Title: TO_AITION

Creator: Sander W. van der Laan

Principal Investigator: Sander W. van der Laan, Imo Hoefer, Gerard Pasterkamp

Data Manager: Saskia Haitjema, Sander W. van der Laan

Project Administrator: Sander W. van der Laan

Affiliation: Other

Funder: European Commission

Template: UMC Utrecht DMP

ORCID iD: 0000-0001-6888-1404

ORCID iD: 0000-0002-8379-7814

ORCID iD: 0000-0001-5345-1022

Project abstract:

Cardiovascular diseases (CVD) such as myocardial infarction and stroke constitute the main cause of morbidity and mortality worldwide, accounting for 3.9 million deaths in Europe (45% of all deaths) and 17.9 deaths globally per year (WHO 2018; EHN 2017). Traditional risk factors-comorbidities of CVD have been obesity, diabetes and hypertension, promoting disease development and severity. Yet, it has been known for decades that an intimate relationship between CVD and depression also exists (Moussavi et al. Lancet 2007). Several studies have demonstrated, that one in three patients with CVD suffer from depression, and depression increases the likelihood for cardiac morbidity and mortality in these patients by 2-3-fold, independently of traditional risk factors or gender (Pelletier BMJ Open 2015). Women with coronary heart disease (CHD) are twice more likely than men to suffer from depression, and women younger than 55 who have depression and premature CHD are at the highest risk for poor outcomes including death compared to younger men and older women, indicating an important age and gender dimension (Gan Y et al. BMC Psychiatry 2014). This relationship is graded as the more severe the depression is, the higher the subsequent risk of mortality and other cardiovascular events. This relationship is also bidirectional as people suffering from depression but do not have CVD have a greater risk of developing heart disease and having a heart attack (Nicholson A. Eur Heart J 2006). CVD and depression develop at the same at-risk population, genetically predisposed individuals that are overweight, smoke, have high cholesterol levels and/or hypertension. They also share a common pathophysiological characteristic, chronic low grade systemic inflammation marked by the prolonged activation of the innate immune system, macrophages, NLRP3 inflammasome and Toll-like receptors, and the expression of pro-inflammatory cytokines such as IL-1, TNF and IL-6 which can

directly cause diabetes, cardiovascular disease or depression (Ferrucci and Fabbri. Nat Rev Cardiol 2018). This raises the possibility that the magnitude, persistence and molecular characteristics of systemic inflammation are the deciding factors determining whether an individual will develop isolated, comorbid or multimorbid disease. Still, the inflammatory state of these diseases has not been systematically analyzed, while the immune-metabolic pathways that are associated and/or causally linked to co/multimorbidity have not been dissected. Moreover, the contribution of socioeconomic, lifestyle and environmental factors including the diet and the gut microbiome, or other metabolic and mental comorbidities, to this process has not been investigated. It remains, therefore, unclear which individuals at-risk are likely to develop isolated, comorbid or multimorbid disease, which treatments work best, and which molecules and pathways constitute plausible targets for the development of novel therapeutics targeting common mechanisms and exhibiting multiple beneficial effects. Through the analysis of TO_AITION's cohorts, the trajectory of multimorbid disease development will further be characterized, new biomarkers identified, and a new diagnostic test, in the form of a lab-on-chip, and risk prediction tool developed. TO_AITION will thus build on these strong foundations in terms of clinical cohorts, omics datasets and immunology to unravel the common causative mechanisms driving comorbid or multimorbid manifestation of CVD-depression and other metabolic or mental disorders in order to advance prevention, diagnosis, prognosis, therapy and management of these highly prevalent and devastating conditions.

ID: 72863

Start date: 01-01-2021

End date: 31-12-2024

Last modified: 20-09-2021

Copyright information:

TO_AITION

1. General features

1.1. Please fill in the table below. When not applicable (yet), please fill in N/A.

DMP template version	29 (don't change)
ABR number *(only for human-related research)*	n/a
METC number *(only for human-related research)*	C-01.18
DEC number *(only for animal-related research)*	n/a
Acronym/short study title	TO_AITION
Name Research Folder	
Name Division	Laboratories, Pharmacy, and Biomedical genetics
Name Department	Central Diagnostics Laboratory
Partner Organization	
Start date study	2021-01-01
Planned end date study	2024-12-31
Name of datamanager consulted*	Saskia Haitjema
Check date by datamanager	

1.2 Select the specifics that are applicable for your research.

- Fundamental / translational study
- WMO
- Use of Questionnaires
- Observational study
- Prospective study
- Retrospective study
- Multicenter study

We will use data from the Athero-Express Biobank Study, which started in 2002 and is ongoing.

2. Data Collection

2.1 Give a short description of the research data.

Subjects	Volume	Data Source	Data Capture Tool	File Type	Format	Storage space
Human	1	Genotype data	R, SNPTEST, GCTA, etc	PLINK-format, Oxford-format	.vcf, .bed/.bim/.fam, .gen/.sample	3Tb
Human	1	RNAseq	R	Binary Alignment Map	.bam	±1Tb
Human	1	scRNAseq	R	Binary Alignment Map	.bam	±1Tb
Human	1	DNA methylation	R	IDAT	.idat	±1Tb
Human	1	OLINK	R	Comma separated	.csv	0-1Gb

2.2 Do you reuse existing data?

- Yes, please specify

Existing data from the Athero-Express Biobank Study:
- clinical, questionnaire data
- histological data
- RNAseq data
- DNA methylation data
- scRNAseq data
- Genotype data

2.3 Describe who will have access to which data during your study.

Type of data	Who has access
Direct identifying personal data	Clinician involved, Datamanager
Key table linking study specific IDs to Patient IDs	Clinicians involved, Datamanager
Pseudonymized data	Research team, Datamanager

2.4 Describe how you will take care of good data quality.

#	Question	Yes	No	N/A
1.	Do you use a certified Data Capture Tool or Electronic Lab Notebook?	x		
2.	Have you built in skips and validation checks?	x		
3.	Do you perform repeated measurements?	x		
4.	Are your devices calibrated?	x		
5.	Are your data (partially) checked by others (4 eyes principle)?	x		
6.	Are your data fully up to date?	x		
7.	Do you lock your raw data (frozen dataset)	x		
8.	Do you keep a logging (audit trail) of all changes?	x		
9.	Do you have a policy for handling missing data?	x		
10.	Do you have a policy for handling outliers?	x		

2.5 Specify data management costs and how you plan to cover these costs.

#	Type of costs	**Division** ("overhead")	Funder	**Other** (specify)
1.	Archiving	x		
2.	Storage	x		
3.	Maintenance Athero-Express		x	
4.	Datamanager	x		
5.	Data analysis tool		x	

2.6 State how ownership of the data and intellectual property rights (IPR) to the data will be managed, and which agreements will be or are made.

UMC Utrecht is and remains the owner of all collected data for this study. There is a consortium agreement which describes the rules and regulations within the consortium regarding IPR. In principle IPR, when applied for and when data from the UMC Utrecht is involved, will be from UMC Utrecht.

3. Personal data (Data Protection Impact Assessment (DPIA) light)

Will you be using personal data (direct or indirect identifying) from the Electronic Patient Dossier (EPD), DNA, body material, images or any other form of personal data?

- Yes, go to next question

3.1 Describe which personal data you are collecting and why you need them.

Which personal data?	Why?
Clinical and questionnaire data, e.g. age, gender (sex), body measurements, medical history, etc.	To answer the research question.
Biomaterial (blood and plaque to isolate DNA, RNA, proteins)	To answer the research question.

3.2 What legal right do you have to process personal data?

- Study-specific informed consent

3.3 Describe how you manage your data to comply to the rights of study participants.

Right	Answer
Right of Access	Research data are coded, but can be linked back to personal data, so we can generate a personal record at the moment the person requires that. This needs to be done by an authorized person.
Right of Rectification	The authorized person will give the code for which data have to be rectified.
Right of Objection	We use informed consents.
Right to be Forgotten	In the informed consent we state that the study participant can stop taking part in the research. Removal of collected data from the research database cannot be granted because this would result in a research bias.

3.4 Describe the tools and procedures that you use to ensure that only authorized persons have access to personal data.

We use a secured Research Folder Structure that ensures that only authorized personnel has access to personal data, including the key table that links personal data to the pseudoID.

3.5 Describe how you ensure secure transport of personal data and what contracts are in place for doing that.

In case we need to transport personal data with colleagues, we use Surffilesender with encryption.

In such events that we collaborate with outside collaborators, we first set up a Research Agreement and/or Data Transfer Agreement regardless of the current consortium agreement.

4. Data Storage and Backup

4.1 Describe where you will store your data and documentation during the research.

The digital files will be stored in a secured Research Folder Structure of the UMC Utrecht. We will need +/- 5 Tb storage space, so the capacity of the network drive will be sufficient.

For purposes of analyses digital files are partly and temporarily stored on the high-performance computer cluster (HPC) facilitated by the institute or a UMC Utrecht owned and managed device.

Paper dossiers are stored safely in a locked cabinet in a locked room in the UMC Utrecht.

Data storage is only accessible to authorized personnel.

4.2 Describe your backup strategy or the automated backup strategy of your storage locations.

All (research) data is stored on UMC Utrecht networked drives from which backups are made automatically twice a day by the division IT (dIT).

5. Metadata and Documentation

5.1 Describe the metadata that you will collect and which standards you use.

For the clinical and questionnaire data collected, we prepared a codebook of the research database. We do not use metadata standards yet.

For the 'omics' data (Protein-, RNA-, DNA-derived) We are collecting metadata in the format of the DUBLIN CORE metadata standard for all our output to enable the smooth publishing of the data in a public repository like Zenodo later. For domain specific data we are currently considering applying the Minimum Information About a Microarray Experiment (MIAME) for our RNA/DNA data.

5.2 Describe your version control and file naming standards.

We will use GitHub as version control for our code based on an existing template (AE_TEMPLATE). There will be a specific GitHub repository for the quality control and analysis of the newly obtained data, in this case protein measurements derived from the OLINK-platform.

We will use the release-system native of GitHub and where possible link it to Zenodo.

6. Data Analysis

6 Describe how you will make the data analysis procedure insightful for peers.

We will write an analysis plan in which we state why we will use which data and which statistical analysis we plan to do in which software. The analysis plan will be stored at GitHub or potentially through a pre-registration server, e.g. OSF. This way this will be findable for our peers.

7. Data Preservation and Archiving

7.1 Describe which data and documents are needed to reproduce your findings.

The data package will contain: the raw data, the study protocol describing the methods and materials, the script to process the data, the scripts leading to tables and figures in the publication, a codebook with explanations on the variable names, and a ‘read_me.txt’ file with an overview of files included and their content and use.

Where it is relevant this is amended by an Electronic Lab Notebook (ELN) and handwritten (legacy) lab journals.

After finishing the project, documentation for the ELN will be stored at the UMC Utrecht [GIVE FULL PATH] and is under the responsibility of the Principal Investigator of the research group.

** I will update ‘XXX’ in this answer when available.*

7.2 Describe for how long the data and documents needed for reproducibility will be available.

Data and documentation needed to reproduce findings from this WMO study will be stored for at least 10 years.

7.3 Describe which archive or repository (include the link!) you will use for long-term archiving of your data and whether the repository is certified.

We will use Archivemetica and DataverseNL to archive and post data according to the principles of FAIR. At the same time a copy will remain at the department server in the existing Research Folder Structure and is under the responsibility of the Principal Investigator of the research group.

7.4 Give the Persistent Identifier (PID) that you will use as a permanent link to your published dataset.

When we get DOI-codes we will update this plan to included these.

8. Data Sharing Statement

8.1 Describe what reuse of your research data you intend or foresee, and what audience will be interested in your data.

1. My peers will be reusing all research data in the final dataset to generate new research questions.
2. The raw data can be of interest for other researchers or for spin off projects.
3. Our processed genetic data can be of interest for other Europeans researchers in the field.

8.2 Are there any reasons to make part of the data NOT publicly available or to restrict access to the data once made publicly available?

- Yes (please specify)

As the data is privacy-sensitive, we publish the descriptive metadata in the data repository with a description of how a data request can be made (by sending an email to the corresponding author). In the event that peers like to reuse our data this can only be granted if the research question is in line with the original informed consent signed by the study participants. Every application therefore will be screened upon this requirement. If granted, a data usage agreement is signed by the receiving party.

8.3 Describe which metadata will be available with the data and what methods or software tools are needed to reuse the data.

The publication will be open assessable. The study protocol and this Data Management Plan will also be available.

Along with the publication, the codebook of the data and scripts of analyses will be available through GitHub.

Data (raw or processed) will be accessible but under conditions and following (local) rules and regulations, and only after obtaining a Data Sharing Agreement.

In the event we share data via the European Genome and Phenome Archive (EGA) is complimented with the limitations of the respective informed consent given by the human participant of the original study/research project. An overview of the allowed broad consent range per individual for each data-file will be provided together with the data collection.

8.4 Describe when and for how long the (meta)data will be available for reuse

- (Meta)data will be available as soon as article is published

8.5 Describe where you will make your data findable and available to others.

We will publish and archive our (sc)RNA-seq, DNA methylation, and genotype data via a suitable repository, for instance the European Genome and Phenome Archive (EGA) or DataverseNL. A minimum set of metadata is publicly available and we will refer to this data-set in our related publication. Interested parties can request access to this data-set via our institutional data access committee (DAC); based on the assessment of the DAC access to this data-set will be granted or declined. When access is granted a Data Sharing Agreement is setup and signed.