Plan Overview

A Data Management Plan created using DMPonline

Title: GoNL

Creator: Sander W. van der Laan

Principal Investigator: Jessica van Setten, Kristel van Eijk, Charlotte N. Onland-Moret, Sander W. van der Laan, Vinicius Tragante do Ó

Data Manager: Jessica van Setten, Kristel van Eijk, Charlotte N. Onland-Moret, Sander W. van der Laan, Vinicius Tragante do Ó

Project Administrator: Jessica van Setten, Kristel van Eijk, Charlotte N. Onland-Moret, Sander W. van der Laan, Vinicius Tragante do Ó

Affiliation: Other

Funder: European Commission

Template: UMC Utrecht DMP

ORCID iD: 0000-0002-4934-7510

ORCID iD: 0000-0002-2360-913X

ORCID iD: 0000-0001-6888-1404

ORCID iD: 0000-0002-8223-8957

ID: 80368

Start date: 01-01-2021

End date: 01-01-3000

Last modified: 20-04-2022

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

GoNL

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)*	n/a
DEC number *(only for animal-related research)*	n/a
Acronym/short study title	GoNL
Name Research Folder	smb://ds.umcutrecht.nl/data/LAB/lab_research/RES-Folder-LKCH/GoNL
Name Division	Laboratories, Pharmacy, and Biomedical genetics
Name Department	Central Diagnostic Laboratory
Partner Organization	
Start date study	2021-01-01
Planned end date study	3000-01-01
Name of datamanager consulted*	Saskia Haitjema
Check date by datamanager	January 6th 2022

1.2 Select the specifics that are applicable for your research.

- Fundamental / translational study
- Non-WMO

We will use data from the GoNL (https://www.nlgenome.nl) as reference in many studies or as part of a course curriculum in practicals to learn genetic analyses methods.

The Genome of the Netherlands is a consortium of the UMCG,LUMC, Erasmus MC, VU university and UMCU, led by Professor Cisca Wijmenga. Samples where contributed by LifeLines, The Leiden Longevity Study, The Netherlands Twin Registry (NTR), The Rotterdam studies, and The Genetic Research in Isolated Populations program. All the sequencing work is done byBGI Hong Kong. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL).

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	±1Tb

2.2 Do you reuse existing data?

- Yes, please specify

Existing data from the GoNL:
- Genotype data
- Some 'clinical' data, i.e. age (when available), sex, relationships (parent-child)

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

Please note, that the data has been de-identified for the purpose of public sharing.

Type of data	Who has access
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 Dataset		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.

The Genome of the Netherlands data distributed on this page is publicly available and can be used under the condition of citation (main paper, additional papers when appropriate).

- The Genome of the Netherlands Consortium. **Whole-genome sequence variation, population structure and demographic history of the Dutch population.** Nature Genetics (2014) doi:10.1038/ng.3021. See (pubmed)
- Francioli *et al.* **Genome-wide patterns and properties of de novo mutations in humans** Nature Genetics (2015) doi:10.1038/ng.3292. See (pubmed)

In addition, the following acknowledgements can be used when reusing public data on a website:
"This study/database makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from http://www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI)."

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?

- No, go to 4.1

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 +/- 1 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.
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).
We will have multiple copies 1) at the HPC, and 2) at the UMC internal network.

5. Metadata and Documentation

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

We do not collect anything else, but the data we can obtain through a download. This includes relationships (parent-child), age (when available), and sex.

5.2 Describe your version control and file naming standards.

We will use GitHub as version control with a specific GitHub repository for the each individual project.
We will use the release-system native of GitHub and where possible link it to Zenodo (code only!).

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 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.
After finishing the projects, documentation will be stored at the UMC Utrecht under the responsibility of the Principal Investigator of the research group and at their respective RFS.
GoNL maintains the data, see section 1.2.

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 do not 'own' the data, it is controlled/managed by the Genome of The Netherlands Project (https://www.nlgenome.nl/menu/main/home). We will only keep copies for local use, and potentially archive projects through Archivemetica and share codes used publications etc through DataverseNL 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.

Specifically the methods and codes developed for the use of this data will be of interest to our peers. Since the data is managed by the GoNL Consortium we refrain from stating anything regarding data re-use, other than that in general these data make for an excellent population reference for multiple purposes.

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, and managed by the GoNL Consortium we will refrain from sharing these data publicly. This should go via the GoNL Consortium.

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

Publications will be open access. 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 under conditions set forward by the GoNL Consortium.

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

- Other (please specify)

Meta data will be accessible under conditions set forward by theGoNL Consortium.

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

We will publish and archive publication, codes, etc as described above through Archivemetica (local archiving) and DataverseNL (public) with a note that the data will be accessible under conditions set forward by the GoNL Consortium.