Plan Overview

A Data Management Plan created using DMPonline

Title: Next generation immunodermatology (NGID)

Creator: Anna Niehues

Principal Investigator: Robert Rissmann

Data Manager: Tom Ederveen

Affiliation: Leiden University Medical Center

Funder: Netherlands Organisation for Scientific Research (NWO)

Template: Data Management Plan NWO (September 2020)

ORCID iD: 0000-0002-5867-9090

Project abstract:

NGID in a nutshell: NGID involves the entire wide range of disciplines required to revolutionize dermatological care. The connectivity of NGID and its perfect momentum in the current era of high-end data acquisition and availability in computing power, is illustrated by the unique blend of researchers that join forces in our consortium: technologists provide the methodology to characterize immuno-dermatological diseases far beyond the current state-of-the-art, biologists fuel their translational models with this information to precisely study disease mechanisms/pathways, leading to the development of novel drugs or tailor-made treatments by pharmacologists and clinicians to perform prospective clinical trials. These insights are then applied by dermatologists in daily clinical practice with the aid of the patient's perspective on their disease and proposed treatment. The important role of psychosocial factors on the effect of treatment and vice versa is safeguarded by psychologists. Internal and external inclusive communication, dissemination and implementation are guided by communication experts. In short, dermatologists base their treatment/care on insights from biologists and analyses developed by technologists, while taking into account the patient needs and psychosocial profile.

ID: 104864

Start date: 01-07-2023

End date: 01-07-2029

Last modified: 26-09-2023

Copyright information:

Next generation immunodermatology (NGID)

General Information

Name applicant and project number

Main applicant: prof. dr. R. Rissmann
Project number : NWA1389.20.182

Name of data management support staff consulted during the preparation of this plan and date of consultation.

Radboudumc Technology Center Data Stewardship

1. What data will be collected or produced, and what existing data will be re-used?

1.1 Will you re-use existing data for this research?

If yes: explain which existing data you will re-use and under which terms of use.

- Yes

NGID will be using data from 4 deep-phenotyping clinical trial in psoriasis (SKINERGY PP) , cutaneous T-cell lymphoma (SKINERGY CTCL), cutaneous lupus erythematosus (SKINERGY CLE) and chronic urticaria (SKINERGY CU) from the co-funder: CHDR (see letter of commitment). The data will be shared with dedicated agreements (consortium agreement) and will only become available in 2023/2024 after startup of the NGID Project.
Authorization to the data will be granted by CHDR upon meeting certain criteria upon data access request.

1.2 If new data will be produced: describe the data you expect your research will generate and the format and volumes to be collected or produced.

Summary of NGID research data to be produced (data format and volume) as described in the approved NGID proposal, break-down based on each relevant Project Task :

Task 1.1.1. Data on the specifications and requirements of the data infrastructure and specific components. User needs are translated into IT parameters, such as: (meta)data collection, standardization, FAIRification, storage, access control, and analysis, and legal and ethical requirements (GDPR).
***Goal**: To determine the requirements of the data infrastructure based on user needs.*

***Sample size expected**: NA*

***Data volume**: questionnaire text only, very minor data volume*

***Notes**: This will be also specified and updated in this Data Management Plan, together with descriptions of the specific (meta)data and possible storage at the partners' sites, including regulatory, operational, storage, and computational needs.*

NB in Task 1.2.1. SOPs will also be developed, but for skin sample harvesting, processing and storage.

Task 1.1.3. Data on definitions of interoperability standards, SOPs and guidelines for omics (meta)data standardization and FAIRification.

***Goal**: To facilitate data integration across partners and modalities as well as supporting future data re-use.*

***Sample size expected**: depends on number of datasets and -types described here*

***Data volume**: SOPs and (FAIR) guidelines only, very minor data volume*

***Notes**: This includes SOPs for storage, retrieval and exchange of data. Each type of (omics)data produced within this project will follow the respective community-standard format(s). In order to be made FAIR, it will be accompanied by rich human- and machine-understandable (ontology-supported) metadata, by using unique persistent metadata. For definition of a data model, we will build upon guidelines and tools from previous projects to implement FAIR principles and thereby facilitate all (multi-omics) analysis. Experimental metadata will be stored in the ISA (short for Investigation, Study, Assay) format to capture provenance of samples and (processed) data.*

Task 2.1.1. Self-report questionnaires with information on:

- distress (HADS);
- perceived stigmatization (6-item stigmatization scale);
- treatment expectancies (custom scales);
- scratching behaviour (ISDL scratching subscale);
- treatment adherence (e-diary);
- Quality of Life (skindex, DLQI);
- treatment satisfaction (TSQM).

***Goal**: To develop patient-centered screening tools to assess psychological/behavioural markers of disease outcomes.*

***Sample size expected**:*

***Data volume**: questionnaire text only, very minor data volume*

Task 2.1.3. Data resulting from conducting joint project meetings, focus groups and semi-structured interviews with patient representatives.

***Goal**: To ensure that the patients' wishes, needs and preferences are prioritized.*

Sample size expected**: **TBD

***Data volume**: report and interview text only, very minor data volume*

Task 2.1.5. Data related to the analysis of the communicative aspects within NGID, by a stepwise, iterative process of feedback which is continuously fed back into infrastructure development (WP1.1) to optimize the prototype.

***Goal**: To assess the needs, preferences, and experiences of relevant stakeholders to ensure that the NGID-infrastructure is optimized for successful implementation.*

***Sample size expected**: NA*

***Data volume**: report and interview text only, very minor data volume*

Task 2.2. Clinical data generation, from observational trials for each of the six dermatological diseases as part of NGID:

- psoriasis (N=120 patients);
- lupus (N=120);
- atopic dermatitis (N=120);
- mycosis fungoides (N=120);
- urticaria (N=120);
- hidradenitis suppurativa (N=120).

Goal*: Trials will be performed with an observational, and an interventional part, from which validated biomarkers will be obtained. These trials are conducted per disease (with varying severities) in different centralized locations.*

Data volume*: very large, many different (omics)data types*
Sample size expected: 100-120 patients for each disease type, 2-3 treatment arms per disease = N=40 for each treatment arm per disease
Approx. 1700 biopsies; 2500 swabs, 4000 tubes
Data volume*: very large, many different (omics)data types*
Metabolomics (plasma): ca. 1700 samples
Proteomics (plasma): ca. 1700 samples
Transcriptomics (RNAseq- biopsies): 1700 biopsies
Spatial metabolomics: ca. 120 samples (substudy)
Tissue CyTOF: ca. 150 samples
Microbiomics: ca. 1500 samples
Pictures: ca. 10000
Note that this basically is part of Task 1.1.1.

1.3. How much data storage will your project require in total?

- >1000 GB

2. What metadata and documentation will accompany the data?

2.1 Indicate what documentation will accompany the data.

Documentation will be part of the general NGID FAIRification processes, meaning that all steps and details from patient including to sampling, processing and data generation will be reported and documented.
Apart from that, for generation of (omics)data and downstream bioinformatics we will maintain our code, SOPs and other guidelines in our GitLab system, which has readme files and Wiki-like documentations for each specific (project) task and application.
As for the clinical trial(s) - all data processes will be performed to SOPs which includes a thorough documentation of all processes and documentation of all particularities. The data will be captured in an electronic Case Report Form (eCRF). All data that will be captured from the tissue will be

documented in specific data report where specifications and deviations will be captured.

2.2 Indicate which metadata will be provided to help others identify and discover the data.

The metadata that we provide is too elaborate to detail here, but we will make use of FAIR community standards depending on the type of omics data wherever available, or build project-specific standards ourselves where needed, and publish these after the project.

For example, metabolomics, proteomics, transcriptomics, microbiomics and imaging data all have their separate requirements in terms of metadata annotations. For many of these we have FAIR templates available, for example provided by resources such as MetaboLights, FAIR genomes, PhenoPacket, ENA, EMBL-EBI, X-omics, MGnify, ELIXIR, etc.

Furthermore, where relevant and applicable we will make use of the ISA format for Investigation, Study, and Assay, to document in FAIR manner our study designs including generated (meta)data and its annotations.

3. How will data and metadata be stored and backed up during the research?

3.1 Describe where the data and metadata will be stored and backed up during the project.

- Institution networked research storage

We will have multiple different forms of (meta)data storage, documentation and archiving. These depends a.o. on the nature of the type of data, its sensitivity in terms of privacy i.e. GDPR, and also on computational requirements for storage and analysis.

Main resources available to us are:

- Institutional networked research storage;
- Digital Research Environment (DRE);
- FAIR Data Points (FDP);
- GitLab systems implemented within the institutes;
- Castor electronic clinical data management platform;

3.2 How will data security and protection of sensitive data be taken care of during the research?

- Default security measures of the institution networked research storage

Yes, we will mainly use default security measures of the institutional network research storage (i.e. servers and data cannot be directly accessed from outside of the hospital networks). In addition, for resources such as the DRE and FDP we can restrict access to sensitive data for specific groups or individuals. This is not only to ensure patient privacy protection wherever needed, but also to facilitate and control data ownership before sharing pseudo- or anonymised research data with the public.

4. How will you handle issues regarding the processing of personal information and intellectual property rights and ownership?

4.1 Will you process and/or store personal data during your project?

If yes, how will compliance with legislation and (institutional) regulation on personal data be ensured?

- Yes

For this we refer again, in part, to 3.1 in this DMP.
In addition, we are well aware that we are by law bound to adhere to the GDPR data protection legislation, and have systems in place to ensure this. Many of the partners within NGID already operate from within academic research centra, and therefore will automatically make use of systems such as EPIC, which can only be visited by trained medical personele.
Any data derived from patient data is pseudo- and where possible anonymised before it will be distributed to (still) highly protected and encrypted systems as described earlier (e.g. Castor, DRE, FDP, etc.).

4.2 How will ownership of the data and intellectual property rights to the data be managed?

These details are described in the final consortium agreement, which defines data and topline management thereof. Furthermore, intellectual property rights were agreed upon in the CA as well.

5. How and when will data be shared and preserved for the long term?

5.1 How will data be selected for long-term preservation?

- All data resulting from the project will be preserved for at least 10 years

We will do as much as possible to ensure data preservation for relevant and valuable research data as long as possible, for instance through archiving in appropriate databases specific to the type of (omics)data, and/or through FAIRification of our data and distributing metadata thereof in our FDPs and platforms such as Zenodo.GCP requires to store data of clinical trials for at least 25 years, the consortium together with the university medical centers and their archiving system will comply accordingly.

5.2 Are there any (legal, IP, privacy related, security related) reasons to restrict access to the data once made publicly available, to limit which data will be made publicly available, or to not make part of the data publicly available?

If yes, please explain.

- Yes

Non-anonymized and non-pseudonymized data as well as pictures etc need to be approved to be used by the participating patients and trial subjects. Only after their written approval in the informed consent form this data can be used for publication and can made publicly available.

5.3 What data will be made available for re-use?

- All data resulting from the project will be made available

see 5.1

5.4 When will the data be available for re-use, and for how long will the data be available?

- Data available after completion of project (with embargo)

[I'm not sure, I assume this will be decided together and described in the CA ?]

5.5 In which repository will the data be archived and made available for re-use, and under which license?

see 5.1
Note that licences will also depend on the databases and scientific journals in which the data will be published and reported.

5.6 Describe your strategy for publishing the analysis software that will be generated in this project.

Any analysis software used and developed during or as part of the project will be sufficiently referenced and published in appropriate Git version-control systems such as GitLab, GitHub, DockerHub, DRE, etc. Likewise, software notebooks (e.g. Jupyter, R-Markdown) and workflow management systems such as NextFlow or Python Snakemake which contain analysis code and workflows will be made available in the appropriate platforms for these.

6. Data management costs

6.1 What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

These resources (and FAIR strategies) are extensively described in the original project proposal, see WP01 *Title: Data infrastructure*, among others, which is a work package which is entirely dedicated to

data management and FAIRification.