ZonMw guidance  A description or narrative of your data collection

A description or narrative of your data collection and what you plan to do, helps to define what is needed for data management.

The DMP template of the UMCG (Groningen) (step 2: Describing your data) shows examples for various scenarios of quantitative and qualitative research:

Describing your data

In this step you describe the data that will be collected, generated or used. This can be newly collected data or (partial) re-use of data.

Please consider the following questions in your answer. Not all question may apply to your research:

- What type of data will be collected, processed, analysed, and/or stored?
- How will the data be created/collected? Which software, tool and methodology?
- Is the data newly collected? And why? Did you check whether the data is already available/accessible?
- Is pre-existing data being used and if so from where?
- What is the format of the datasets?
- How is version control of the data arranged? How do you ensure you use the right version?
- What is considered ‘raw’ data in your study and what is processed data?
- How do you obtain processed or cleaned data?
- How is the data quality and data processing is controlled?
- In case of data from human subjects: how do you pseudonymise or anonymise your data?
- Does your study collect/process/analyse personal and/or direct (e.g. name) or indirect identifiable data (use of a combination of variables like birth data, gender, geographical location and a disproportionate investment of time and effort)? Why do you need this data?

EXAMPLES

Newly collected questionnaire data

We will newly collect data on demographic information and on quality of life via a digital questionnaires, using tool XYZ for 100 patients at the dermatology outpatient department. Patient details such as disease history will be collected using paper questionnaires. In addition we will measure xxx function, using equipment XYZ. No data from medical records will be collected for this study. Paper questionnaires will be digitized and data entered into a regular database format (type XYZ). Questionnaires and patient details will be collected using an eCRF-system suited for the data collection < e.g. Open Clinica, Oracle Clinical, Utopia, Roqua, etc>. These raw datafiles will be locked for editing after closure of the study. Data cleaning and processing will be executed using syntaxes in programme XYZ. Data analyses will take place on the cleaned data that does not contain the (in)direct identifiable data, using programme XYZ. All the syntaxes describing data processing and analyses will contain or be accompanied by the comments explaining the code and the decisions made. The syntaxes will be checked by a data quality controller.

We will minimise the collection of possible identifiable information; names and addresses will not be collected. All participants will receive a research ID number in the dataset, using a subject
identification code. The Subject identification code list will be maintained by the investigator. Year of birth, gender and geographical location (postal code 4) are necessary for the research. These data are newly collected, because it is not yet available. All data can be relevant for future re-use by internal and external colleagues.

Re-use of genetic data
For this study we will combine next generation sequencing data from three existing biobanks, of in total 600 subjects. To obtain the data a data request procedure has been followed and finalised. A data access agreement is put in place for all three biobanks. The raw data (format XYX) will be processed using specifically designed bioinformatics pipelines on a High Performance Computing (HPC) cluster in a Linux environment creating quality controlled data files (format XYX). These files will be used for data analysis. All scripts are R based, tested and stored with the research documentation. The data files do not contain direct identifiable data; all subjects have a subject ID. For each biobank a small set of phenotypic data on disease outcome is available, that can be linked to the genetic data using the subject ID.
This study uses re-useable data. Other researchers can contact the biobanks for re-use of the data.

Qualitative research using video’s
For this research on mental health, 20 children (<18 year) will be interviewed. In addition, we will use existing data on education attainment using the XYZ dataset, that is publically available/data access has been requested. We have permission of the owner of the data for re-use. The interviews will be recorded using video tapes to capture non-verbal communication. The tapes are transcribed by company XYZ. The transcripts in word documents will be coded (using XYZ) and transformed to a databases format. These files will be combined with the education attainment dataset the using a Trusted Third Party.
The video’s contain identifiable information, but the transcribed and coded dataset that are used for data processing and analysis not. Personal data is only used for linkage with the education attainment dataset and after linkage removed. The combined dataset does not contain personal data. We aim to re-use data where possible.
The interview data is not yet available from other studies. We searched database XYZ and catalogues XYX for this. All data can be relevant for future re-use for internal and external colleagues.

(Neuro)imaging study
In this research project we will collect and analyse imaging data from 100 retrospectively selected patients who underwent thoracic CT for tumor assessment. This data collection will be evaluated as a non-WMO study and all patient numbers will be checked against the no-objections registration (geen-bezwaar registratie). Only the original DICOM series will be selected, all derived data such as Secondary Captures (SC), Structured Reports (SR), etc. will be excluded from the selection to avoid identity breach since they may contain identifiable information even after de-identification of the DICOM header.
The collected data will be stored in DICOM format. The user-identifiable information from the DICOM header will be removed or de-identified upon extraction from the institutional Picture Archive and
Communication System (PACS). Further analysis of the data will be performed using software XYZ and results will be recorded in OpenClinica.

NOTE: With imaging data, some additional caveats should be considered:
1) If you use captured data, e.g. Ultrasound, patient identification information needs to be blanked out from the image itself (not only from the header).
2) Be extra careful when dealing with 3D imaging data of the head. Volumetric reconstructions can be made from this data revealing the actual face of the scanned person, thus providing information about the identity.