Horizon DMP example answers

KI Research Data Office

SUMMARY *(dataset reference and name; origin and expected size of the data generated/collected; data types and formats)*

Origin of data:

- Image files will be recorded from a confocal microscope.
- Patient data will be acquired from the Swedish Hip Arthroplasty Register.
- Survey responses will be acquired using the RedCap survey software.
- Respondent data will be acquired in clinical interviews.
- RNA sequencing data will be generated from normal and tumor tissues from patients.
- Existing data will be used for new analysis.

Data format:

- Biomarker Data will be received/imported/saved in a .csv format.
- Questionnaire data will be saved in SAS format.
- Interview responses will be saved in Nvivo .nvp format.
- Survey responses will be exported from RedCap to .csv format.
- Register data will be received in spreadsheet format and will be converted to .tsv format before analysis.
- Sequencing data will be in fastq format.
- Flow cytometry data will be saved in .fcs format.
- Confocal images will be saved in .jpeg format.
- Proteome raw data will be saved in .raw files
- Raw methylation data will be in .idat format.
- Raw genetic variation data will be in .vcf format.

MAKING DATA FINDABLE *(dataset description: metadata, persistent and unique identifiers e.g., DOI)*

A DOI will be assigned to the dataset by the data repository (e.g. SND).

Documentation will include a standardized folder structure, codebooks (metadata about the data), logbooks (metadata about data processing), analysis plans, input and output files from databases and statistical softwares.

All files will be renamed, with date of acquisition and experimental condition, and put into folders. A "read me" file will be generated, explaining the experimental conditions, tissue and cell types.

Survey responses will be curated into the Psych-DS format.

Patient data will be read into SPSS and working files will be named with a version suffix, e.g. v2.

The following metadata will be provided (as Excel file) for each experiment:Experiment number, Condition, Date, Creator, Description, Format

Data will be documented following the MINSEQE standard recomendations (http://fged.org/projects/minseqe/).

Metabolomics data will be documented in accordance with community standards defined by the Metabolomics Standards Initiative

MAKING DATA OPENLY ACCESSIBLE *(which data will be made openly available and if some datasets remain closed, the reasons for not giving access; where the data and associated metadata, documentation and code are deposited (repository?); how the data can be accessed (are relevant software tools/methods provided?)*

Data will be made available upon publication as a supplement to the publication.

Data will be deposited at a repository/database (please provide name) immediately and without embargo.

Metadata will be deposited at SND and be freely searchable after publication, with links to the underlying data.

Information about data and metadata are available for the register X holder.

Only metadata is published openly, underlying data is made available upon request after ensuring compliance with relevant legislation and KI guidelines.

Analysis scripts and other developed code will be uploaded to Github

MAKING DATA INTEROPERABLE *(which standard or field-specific data and metadata vocabularies and methods will be used)*

The format of the data is (provide details) is suitable for the management, long-term preservation, and accessibility of research data and recommended by SND .

We are using the AGLS metadata standard.

INCREASE DATA RE-USE *(what data will remain re-usable and for how long, is embargo foreseen; how the data is licensed; data quality assurance procedures)*

Tools needed:

- A software license for SAS, STATA, Nvivo, etc will be required.
- The spreadsheet data can be read with any software compatible with .csv files.
- Image files can be opened with any software compatible with .jpeg files.

Licenses/Agreements:

- Data Transfer/Processing agreements will be signed prior to any data sharing.
- Data will be deposited at a repository/database (please provide name) immediately and without embargo, using a license (please specify license type, e.g CC-BY).

Data quality:

- Data will be quality-checked at collection/generation by validation against controls or publicly available databases.
- RNA seq data will be quality controlled in terms of sequence quality, sequencing depth, reads duplication rates (clonal reads), alignment quality, nucleotide composition bias, PCR bias, GC bias, rRNA and mitochondria contamination, coverage uniformity. Only high-quality data will be included in the subsequent analysis.
- The register holder assures data quality in terms of completeness and correctness of registration.
- The transcribed interview material will be coded independently by two researchers.
- Images will be inspected for artifacts and the results will be recorded in a spreadsheet file.
- Register data will be quality controlled according to a procedure established in our group (REF).
- Data input is validated at the point of entry (RedCap) which does not permit missing data. Data will be checked to verify that all responses are within the possible range of data values.
- Data will be checked for double entries, completeness, missing data, unreasonable values and linkage between the data sources.

ALLOCATION OF RESOURCES and DATA SECURITY *(estimated costs for making the project data open access and potential value of long-term data preservation; procedures for data backup and recovery; transfer of sensitive data and secure storage in repositories for long term preservation and curation)*

Allocation of resources:

Data management is performed by the PI / a research assistant / a postdoc / a dedicated data manager.

No specific resources are allocated for data management.

Salary of X SEK for a data manager in the group is required.

Access to the departmental server is required. It is expected to cost X SEK.

Data security:

Access to the documentation stored in ELN is restricted to group members.

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

We will use a local server that is backed-up at the institution for data storage. Working datasets, and metadata will be stored on the server. The server name is XXX and the folder where the data is saved is XXX. Link to the server:(if possible)

KI ELN (automatically backed-up) will be used for the documentation of all analyses and results.

Long-term storage will take place at a server with automatic back-up at the Institution. Data will be stored at least 10 years after publication. The data will include raw data and the final data analysis file. Intermediate working files will be deleted.

Sensitive personal data will be handled according to KI:s guidelines (https://staff.ki.se/gdpr).

We only work with pseudonymized data, with the key stored in a safety cabinet located at XXX (please specify location) and to which only XXX have access to (please specify the people that have access to it).

It has been judged that controlled access is not required for these data since the data do not contain personal information.

Patient data is pseudonymized by the clinical collaborator or the register holder, and the code is not accessible to researchers in our research group. The material will arrive to KI coded, and the original code will be saved by the collaborators/register holder.

Survey and clinical data will be anonymized, i.e. all possibility to trace the data back to the study participant has been removed. The data is anonymized when the code key is destroyed and it is no longer possible to directly or indirectly connect a person to the data.