Horizon Europe

KI RDO example answers:

The following answers are provided for inspiration, some of them might be applicable for your project

1. Data Summary

Will you re-use any existing data and what will you re-use it for? State the reasons if re-use of any existing data has been considered but discarded.

We will reuse data from previous experiments to compare/evaluate new data generated in the project.

What types and formats of data will the project generate or re-use?

- Biomarker Data will be saved in a .csv format.
- PCR data will be saved in .csv format
- Questionnaire data will be saved in SAS format.
- Data on prescribing practices before and after pilot trial will be managed in SAS (file format: .sas7bdat) and analyzed in STATA (file format: .dta).
- 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.

What is the purpose of the data generation or re-use and its relation to the objectives of the project?

The data collection in the project will be used to collect research data from pre-clinical and clinical studies with selected vaccine candidates against COVID-19 infections.

What is the expected size of the data that you intend to generate or re-use?

A data volume of 3 TB is expected

What is the origin/provenance of the data, either generated or re-used?

- Image files will be recorded from a confocal microscope.
- RNA sequencing data will be generated from normal and tumor tissues from patients.

- Patient data will be acquired from the Swedish Hip Arthroplasty Register.
- Survey responses will be acquired using the REDCap survey software.
- Measurements of markers of liver and renal function will be collected in the SMART-TRIAL system.
- Respondent data will be acquired in clinical interviews.
- Existing bioinformatics data will be used for new analyses.

To whom might your data be useful ('data utility'), outside your project?

2. FAIR data

2.1. Making data findable, including provisions for metadata

Will data be identified by a persistent identifier?

We plan to make our datasets findable by uploading rich metadata to a searchable resource (a data repository) and having a persistent identifier assigned to the data by the repository. Data will be deposited at a repository/database (please provide name) immediately and without embargo

Will rich metadata be provided to allow discovery? What metadata will be created? What disciplinary or general standards will be followed? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

- Data will be described by rich metadata using standard or specified terminologies:
  - ✓ 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 software
  - ✓ All files will be named according to the 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.
  - ✓ Working files will be clearly labelled 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

- ✓ Study documentation procedures have been developed in consultation with and Karolinska Trial Alliance, KTA). File structure and naming has been adapted from templates provided by the KTA.

Will search keywords be provided in the metadata to optimize the possibility for discovery and then potential re-use?

Search keywords will be provided in the metadata

Will metadata be offered in such a way that it can be harvested and indexed?

- Metadata will be deposited at SND and be freely searchable. There will be links to the underlying data.

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

2.2. Making data accessible

Repository:

Will the data be deposited in a trusted repository?

Yes. Data and metadata will be retrievable by their unique and persistent identifier assigned by the data repository.

Have you explored appropriate arrangements with the identified repository where your data will be deposited?

Does the repository ensure that the data is assigned an identifier? Will the repository resolve the identifier to a digital object?

Yes, a DOI will be assigned to the data by the repository

Data

Will all data be made openly available? If certain datasets cannot be shared (or need to be shared under restricted access conditions), explain why, clearly separating legal and contractual reasons from intentional restrictions. Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if opening their data goes against their legitimate interests or other constraints as per the Grant Agreement.

Datasets that do not contain personal information will be:

- made available upon publication as a supplement to the publication.
- deposited at a repository/database (please provide name) immediately and without embargo.

Datasets containing personal information will be:

- deposited in EGA. **For EGA-deposited data, qualified researchers can obtain the data after signing an agreement that includes privacy protection compliant with GDPR, including adequate data protection and the possibility to withdraw consent.**
- Made available upon request after ensuring compliance with relevant legislation and KI guidelines. Metadata will be published open in a data repository.

If an embargo is applied to give time to publish or seek protection of the intellectual property (e.g. patents), specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Will the data be accessible through a free and standardized access protocol?

If there are restrictions on use, how will access be provided to the data, both during and after the end of the project?

How will the identity of the person accessing the data be ascertained?

Is there a need for a data access committee (e.g. to evaluate/approve access requests to personal/sensitive data)?

Metadata

Will metadata be made openly available and licenced under a public domain dedication CC0, as per the Grant Agreement? If not, please clarify why. Will metadata contain information to enable the user to access the data?

How long will the data remain available and findable? Will metadata be guaranteed to remain available after data is no longer available?

Will documentation or reference about any software be needed to access or read the data be included? Will it be possible to include the relevant software (e.g. in open source code)?

Data and metadata will be retrievable by their unique and persistent identifier assigned by the data repository.

2.3. Making data interoperable

What data and metadata vocabularies, standards, formats or methodologies will you follow to make your data interoperable to allow data exchange and re-use within and across disciplines? Will you follow community-endorsed interoperability best practices? Which ones?

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies? Will you openly publish the generated ontologies or vocabularies to allow reusing, refining or extending them?

Will your data include qualified references to other data (e.g. other data from your project, or datasets from previous research)?

We plan to make our datasets interoperable by using controlled vocabularies, keywords or ontologies where possible.

RNAseqData 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

We will also use file formats that are as open and widely used as possible, which will also facilitate **data exchange between partners.**

2.4. Increase data re-use

How will you provide documentation needed to validate data analysis and facilitate data re-use (e.g. readme files with information on methodology, codebooks, data cleaning, analyses, variable definitions, units of measurement, etc.)?

- Data will be described by rich metadata using standard or specified terminologies:

- ✓ 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 software

- ✓ All files will be named according to the 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.

- ✓ Working files will be clearly labelled 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

- ✓ Study documentation procedures have been developed in consultation with and Karolinska Trial Alliance, KTA). File structure and naming has been adapted from templates provided by the KTA.

Will your data be made freely available in the public domain to permit the widest re-use possible? Will your data be licensed using standard reuse licenses, in line with the obligations set out in the Grant Agreement?

- 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 transfer or processing agreement will be performed in the context of our consortia agreement. If necessary, will be performed between our research group and collaborators for data transfer, previously approved by KI's legal department.
- Researchers that would like to reuse the data from EGA will have to accept the EGA terms and conditions for downloading/using data.

Will the data produced in the project be useable by third parties, in particular after the end of the project?

Will the provenance of the data be thoroughly documented using the appropriate standards?

Yes

Describe all relevant data quality assurance processes.

- 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.
- Mass spectrometry results will be quality-checked for contamination and mass accuracy.
- Register data will be quality controlled according to a procedure established in our group (REF).
- Data will be checked at the point of entry in REDCap or SMART-TRIAL for double entries, completeness, missing data and unreasonable values.
- To assure data quality, the study will be conducted according to the COREQ guidelines for qualitative research.

3. Other research outputs

In addition to the management of data, beneficiaries should also consider and plan for the management of other research outputs that may be generated or re-used throughout their projects. Such outputs can be either digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.).

Beneficiaries should consider which of the questions pertaining to FAIR data above, can apply to the management of other research outputs, and should strive to provide sufficient detail on how their research outputs will be managed and shared, or made available for re-use, in line with the FAIR principles.

- https://www.protocols.io/ will be used to share scientific protocols used in the study
- Analysis scripts and other developed code will be uploaded to Github.
- Genetically modified mice strains that will be generated will be made available via the Jackson Labs

4. Allocation of resources

What will the costs be for making data or other research outputs FAIR in your project (e.g. direct and indirect costs related to storage, archiving, re-use, security, etc.) ?

How will these be covered? Note that costs related to research data/output management are eligible as part of the Horizon Europe grant (if compliant with the Grant Agreement conditions)

Who will be responsible for data management in your project?

How will long term preservation be ensured? Discuss the necessary resources to accomplish this (costs and potential value, who decides and how, what data will be kept and for how long)?

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

5. Data security

What provisions are or will be in place for data security (including data recovery as well as secure storage/archiving and transfer of sensitive data)?

- Access to the documentation stored in ELN/REDCap/Onedrive/KI servers is restricted to group members.
- Data saved in ELN/REDCap/Onedrive/KI servers is backed up.
- Access to data saved in ELN/REDCap/Onedrive/KI servers requires user authentication with password.
- Access to ELN/Onedrive/KI servers is permitted only when on KI premises or by VPN or MFA
- The data in ELN/REDCap/KI servers is saved locally at KI. For ELN/REDCap, two redundant servers are used that have standardized physical security.
- All network traffic to and from ELN/REDCap is encrypted.
- For ELN/REDCap/SMART-TRIAL, data access is based on an individual's role in the project.
- ELN/REDCap provide audit trails for tracking data changes and user activity
- In OneDrive, it is possible to recover changed/deleted datasets.
- SMART-TRIAL is only accessible through a secure encrypted web address (Secure Socket Layer (SSL) and Transport Layer Security (TLS) technologies), via a unique user ID and secure password (two-step verification and authentication).
- Human sequencing data from NGI will be processed and temporarily stored in the Bianca server for sensitive data at Uppmax (Uppsala Multidisciplinary Center for Advanced Computational Science), which has several layers of security.
- 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.

Will the data be safely stored in trusted repositories for long term preservation and curation?

The data will be safely stored in repository X for long term preservation and curation

6. Ethics

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Will informed consent for data sharing and long term preservation be included in questionnaires dealing with personal data?

- There are no personal data, nor any other grounds for confidentiality.
- Sensitive personal data will be handled according to GDPR. (https://staff.ki.se/gdpr).
- IP rights will be managed in accordance with the contract drawn up with our industrial partner organization (specify).
- 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 connect a person to the data.
- Data will be pseudonymized and a key will be kept separately from the data.
- Patient data is pseudonymized by the clinical collaborator 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.
- The code key for pseudonymized data is kept by the holders of the original registers, i.e., by the Swedish National Board of Health and Welfare (https://www.socialstyrelsen.se/), Statistics Sweden (https://www.scb.se/), and Region Stockholm (https://www.sll.se/) and not available to us at any time.
- Ethical approvals/amendments and informed consent forms for the project are registered in the diary.
- Consent has been acquired from human participants to process/share data.
- Data Transfer/Processing agreements will be signed prior to any data sharing.
- Results will only be presented on aggregated level without any possibility of backward identification.
- The study will be performed in accordance with the ethical principles of the World Medical Association (WMA) Declaration of Helsinki and aims to follow Good Clinical Practice (GCP) guidelines.

7. Other issues

Do you, or will you, make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones (please list and briefly describe them)?

Institutional policies followed:

- https://staff.ki.se/guidelines-for-research
- https://staff.ki.se/guidelines-for-research-documentation-and-data-management