DESCRIPTION OF DATA

How will data be collected, created or reused?

- 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.

What types of data will be created and/or collected, in terms of data format? Include version numbers if applicable.

- 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.

DOCUMENTATION AND DATA QUALITY

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, file naming-format-versioning, etc

- 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.
- All experimental details will be documented at KI ELN. We will use templates when applicable, which ensures standardized operating procedures.
- 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.

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

- 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.

STORAGE AND BACKUP

How is storage and backup of data and metadata safeguarded during the research process?

- Working datasets, and metadata will be stored on a P folder at a central IT server / a folder at institutional server / OneDrive. Link to the folder (if possible)
- Data saved in ELN/REDCap/Onedrive/KI servers is backed up.
- KI ELN be used for the documentation of all analyses and results.
- Genotyping data will be saved in TicketLab
- During the analysis of the RNA-sequencing data, fastq and analysis files will be stored at the secure cluster Bianca at Uppmax. All files will be transferred to a server at KI when the analysis is over.
- All data in SMART-TRIAL is stored on secured Microsoft Azure hardware located in the European Union

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

- 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.

LEGAL AND ETHICAL ASPECTS

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

- 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).

- Data will be pseudonymized and a key will be kept separately from the data.

- If necessary, data transfer or data processing agreement will be performed between our research group and collaborators for data transfer, previously approved by KI's legal department.

- IP rights will be managed in accordance with the contract drawn up with our industrial partner organization (specify).

How is correct data handling according to ethical aspects safeguarded?

- 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.

- 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.

ACCESSIBILITY AND LONG-TERM STORAGE

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes, licenses and limitations on the access to and reuse of data?

- 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. There will be links to the underlying data.

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

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

- Analysis scripts and other developed code will be uploaded to Github.

- Experimental workflows and protocols will be made available via protocols.io.

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

- Long-term storage will take place at the server at the Institution and in ELN. Data will be stored at least 10 years after publication. The data will include raw data and the final data analysis file.

- As soon as an e-archive is available centrally at KI the data will be transferred to the e-archive.

Will specific systems, software, code or other types of services be necessary in order to open and use/analyse data in the long term?

- The data can be read by any software compatible with .jpeg files
- The data can be read by any software compatible with .csv files
- A software licence for SPSS will be required to read the data file which has been analysed.
- Code necessary to process and interpret the data will be deposited on GitHub.

How will unique and persistent identifiers for the research data, such as a Digital Object Identifier (DOI), be obtained?

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

RESPONSIBILITY AND RESOURCES

Who is responsible for data management while the research project is in progress?

- The researcher who obtains the data also manages the data.
- Data management is performed by a research assistant in the group.
- Data management is performed by a dedicated data manager in the research group, who is an experienced researcher with a PhD.

Who is responsible for data management, long-term storage after the research project has ended?

- The PI is responsible for ensuring that the data is stored safely during and after the completion of the project. The PI is also responsible for contacting the archive at the institution or the central KI archive.

What resources (costs, labour or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)?

- No specific resources are allocated for data management.
- Salary for a data manager in the group is funded X% by this grant.
- Access to the departmental server is required. It is expected to cost X SEK and is covered by the project budget.

What resources will be needed to ensure that data fulfil the FAIR (= Findable, Accessible, Interoperable & Reproducible) principles?

- We will require assistance from the library Data Access Unit to upload the dataset to the SND catalogue.

- 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,

- We plan to make our datasets ACCESSIBLE by ensuring that following the persistent identifier will lead to the data or associated metadata

- We plan to make our datasets INTEROPERABLE by using controlled vocabularies, keywords or ontologies where possible and by using open file formats

- We plan to make our datasets RESUSABLE by assuring high data quality, by providing all documentation needed to support data interpretation and reuse and by clearly licensing the data via the repository so that others know what kinds of reuse are permitted.

- No particular additional resources will be required.