Plan Overview

A Data Management Plan created using DMPonline

Title: Integrative digital diagnostics for therapy optimization in early breast cancer

Creator: Theodoros Foukakis

Principal Investigator: Theodoros Foukakis

Affiliation: Karolinska Institutet

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0000-0001-8952-9987

Project abstract:

Purpose and aims

This project aims to develop tools for prediction of response to neoadjuvant (pre-operative) therapy (NAT) and prognostication of post-surgery risk of recurrence in breast cancer. To this end, input from radiology, digital pathology, genomics and informative clinical variables will be integrated using a **machine learning (ML)-based multi-modal fusion strategy**.

Project organisation, time plan and scientific methods

Three academic clinical trials and one population-based cohort of NAT (N=2500) will be used to train single-source predictive model priors that will be ensembled into integrative multi-omics predictive models. These will be validated externally in independent cohorts of ~3000 patients.

The project will be divided into work packages (WP), corresponding to each of the data modalities. WP1 data and material collection (year 1-4); WP2-3 transcriptomics and genomics in tissue and blood (y 1-3); WP4 radiomics using mammography and magnetic resonance imaging (y 1-3); WP5 pathomics (y 1-3); WP6 model integration (y 3-4); WP7 external validation (y 4-5).

Importance

The project will contribute with novel ML methodology for clinical medicine and a precision oncology solution for optimizing NAT selection and risk stratification that will lead to less over- and under treatment, sparing patients from unnecessary toxicities and reducing financial burden to healthcare systems, and ultimately improving prognosis for patients with breast cancer.

ID: 116982

Start date: 01-01-2022

End date: 31-12-2025

Last modified: 13-10-2025

Grant number / URL: 2021-03061

Integrative digital diagnostics for therapy optimization in early breast cancer

General Information

Project Title

Integrative digital diagnostics for therapy optimization in early breast cancer

Project Leader

Theodoros Foukakis

Registration number/corresponding, date and version of the data management plan

Version 1, 8 Feb 2023

Version

1

Date

8 Feb 2023

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

- Patient and tumor data collected within clinical trials
- Patient data acquired from population-based registries
- Image files (histopathology slides) recorded from Hamamatsu Nanozoomer scanner.
- Image files from GeoMx® Digital Spatial Profiler platform.
- Image files from multiplex fluorescent immunohistochemistry generated from Vectra Polaris.
- DNA and RNA sequencing data generated from fresh frozen tumor samples from patients.
- DNA and RNA sequencing data generated from FFPE (Formalin-Fixed Paraffin-Embedded) tumor material from patients.
- DNA sequencing data generated from patients’ blood/plasma/serum.

- Measurements of tumor biomarkers from patients' plasma/serum.
- Bioinformatics data that have been used or will be used for analyses.
- Radiological data generated from CT scan/MRI used for radiomics-based analysis.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

- Patient and tumor data in .xlsx format.
- DNA and RNA sequencing data in .fastq format.
- Measurements of tumor biomarkers from patients' plasma/serum in .xlsx format.
- Bioinformatics data that have been used or will be used for analyses in .xlsx format.
- Image files (histopathology slides) recorded from Hamamatsu Nanozoomer scanner in .ndpi format.
- Image files from GeoMx® Digital Spatial Profiler platform.
- Image files from multiplex fluorescent immunohistochemistry generated from Vectra Polaris.
- Radiological data generated from CT scan/MRI used for radiomics-based analysis in .dcm 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, 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, etc for each project.

All experimental details will be documented at KI ELN (Electronic Lab Notebook). We will use templates when applicable, which ensures standardized operating procedures.

All data will be accurately described with rich metadata. The metadata will document how the data were generated, under what license and how they can be re-used, and provide the context for proper interpretation by other researchers.

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

Clinical Study documentation procedures have been followed by the clinical trial administration at Center för Kliniska Cancerstudier (CKC), Karolinska University Hospital, in accordance with standard practice and following rules for Good Clinical Practice (GCP) and all applicable legislation. File structure and naming has been adapted from templates provided by the CKC.

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/DNA 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.

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.

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.

Data saved in ELN/Onedrive/KI servers are backed up in external hard drives, stored safely in our group.

KI ELN will be used for the documentation of all analyses and results.

Genotyping data are delivered through Data Delivery System (DDS) which is a cloud-based system for the delivery of data from SciLifeLab. It consists of a command line interface (CLI) and a web interface. (https://delivery.scilifelab.se/)

During the analysis of the DNA/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.

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/Onedrive/KI servers is restricted to group members.

Data saved in ELN/Onedrive/KI servers is backed up in external hard drives, stored safely in our group.

Access to data saved in ELN/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/KI servers is saved locally at KI. For ELN/KI servers, two redundant servers are used that have standardized physical security.

All network traffic to and from ELN /Ki servers is encrypted.

For ELN/KI servers data access is based on an individual’s role in the project.

ELN provides audit trails for tracking data changes and user activity.

In OneDrive, it is possible to recover changed/deleted datasets.

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 at CKC, Karolinska University Hospital for clinical trials and in a safety cabinet located at J5:30, Bioclinicum, NKS for retrospectively collected patient cohorts (eg. Registry data). Only authorized personnel have access to the key for each cohort.

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?

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.

A data processing agreement (Swedish PUB avtal) will be set up between KI and Karolinska University Hospital regarding data from Clinical Trials and other data, for which the hospital is responsible.

If necessary, additional 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.

How is correct data handling according to ethical aspects safeguarded?

Patient data are pseudonymized by our study coordinator and the key-code is not accessible to researchers in our research group (only upon request from the PI with a duty of confidentiality).

Ethical approvals/amendments and informed consent forms for the project are registered in a diary kept in our research group. Consent has been acquired from human participants to process/share data.

If needed, 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.

All studies are performed in accordance with the ethical principles of the World Medical Association (WMA) Declaration of Helsinki and aim 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 and limitations on the access to and reuse of data to be considered?

Data will be made available upon publication as a supplement to the publication, depending on the format and the processing of the data. Raw data are not usually published, but a subset of processed data might be available depending on the publication.

Data will be deposited at a repository/database 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 guielines.

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

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, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

Depending on the format of the data, data can be read by any software compatible with .jpeg files./any software compatible with .csv files.
A software license for SPSS will be required to read the data file which has been analyzed.
Code necessary to process and interpret the data will be deposited on GitHub.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

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

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

Responsible for data management while the research project is in progress, is the researcher who obtains the data, e.g., PhD student/research assistant/postdoc.
Responsible for data management and long-term storage after the research project has ended, is the PI of the group, Theodoros Foukakis. 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 input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

External hard drives for back-up as specified above have been purchased. A fee when the free space in the P server has been exceeded will be paid to KI. Data management is done by members of the group, no data manager will be employed. No other specific resources (costs/labor/other) are allocated for data management (including storage, back-up, provision of access and processing for long-term storage).

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.