DECIDER

Grant Agreement number: 965193

Improved clinical decisions via integrating multiple data levels to overcome chemotherapy resistance in high-grade serous ovarian cancer

Data Management Plan

Main authors: Tiia Pelkonen, Sampsa Hautaniemi

Contributing authors: Ann-Christin Ostwaldt, Diego Boscarino, Olli Carpén, Kaisa Helminen, Johanna Hynninen, Veli-Matti Isoviita, Sanaz Jamalzadeh, Tuula Kallunki, Anna Laury, Yilin Li, Fran Supek, Ole Thastrup

History of changes:

Version	Date	Update(s)
1.0	30.7.2021	First submitted version of the DMP
2.0	30.1.2024	• change of the name “Hospital District of South Western Finland (TUCH)” to ”Wellbeing Services County of South Western Finland (TYKS)” due to a partial takeover of the entity • addition of details on the permanent archiving of sequencing data to the EGA • change of persons responsible for certain areas of the data management • addition of details regarding the storage and usage of CT and PET data
3.0	12.06.2025	• adding additional modes of secure transfer of data between project partners • adding details on CT image analysis

Table of Contents

1. Introduction

Data Management Plans (DMPs) are considered to be a key element to sound data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 Project. The goal of this document is to set the DMP for the DECIDER project. It contains guidelines that will be used by the DECIDER consortium partners with regard to all the data that will be generated by the project.

The DMP will not be a fixed document. This version provides an updated view of the data generated and collected during the project as well as how it will be managed. The DMP will be further updated as needed as the project progresses and will be reviewed and updated by the consortium at every periodic report at the minimum.

Data management here also covers internal documents related to project management. These include research agreements, data and material transfer agreements, data processing agreements, procurement agreements, Non-disclosure agreements (NDAs) and ethical permissions/documentation, which are kept in the project archives as well as in the respective organisational archives of the agreement partners. Project reports (deliverables and periodic reports) are saved in the project's internal documentation, in the EU participant portal project reporting module, and those deliverables with dissemination level 'public' are published on the EU CORDIS project page and the project website.

We have opted to join the Horizon2020 Open Research Data Pilot (ORD pilot) and therefore comply with its core principle to make data "as open as possible, as closed as necessary." Furthermore, we employ the FAIR principle where data are Findable, Accessible, Interoperable and Re-usable. As the goal of DECIDER is to develop tools for personalised medicine, method development and implementation are conducted so that the EU's "right to explanation" principle is met.

2. Data Summary

2.1 Purpose of data collection and generation

The purpose of data collection and generation in the DECIDER project is to characterise drug resistance mechanisms in high-grade serous ovarian cancer (HGSOC) and suggest effective means to overcome them. The project aims to develop a software tool (Oncodash), where all clinically relevant data from a patient can be viewed easily to aid in making treatment decisions. Research data collected and generated in the project will be made available to the wider research community and/or public to the extent that is possible considering ethical, personal data protection and IPR matters.

Patient samples and associated clinical, molecular, and imaging data are collected/generated prospectively before surgery and chemotherapy, between treatments and after disease relapse, and from retrospective samples in order to 1) develop computational tools that pinpoint the drug resistance mechanisms, 2) identify patients who are likely to respond poorly to the current

standard therapy as early as possible, and 3) suggest effective, personalised therapies to patients.

The focus of DECIDER is on the analysis of samples (tissue, ascites, and blood) obtained from HGSOC patients who have given their consent to use their samples and data in the research conducted in DECIDER. This cohort is subsequently called “prospective” because patients are recruited and treated in parallel to the activities in DECIDER. All samples from the prospective cohort are collected during routine procedures conducted in the Turku University Central Hospital (part of the “Wellbeing Services County of Southwest Finland”, TYKS). The key measurement technologies to be used for the samples are various sequencing technologies, in particular whole-genome sequencing (WGS), exome/panel-sequencing (for blood and reference tissue samples only), shallow sequencing (for blood samples only), RNA-seq and DNA methylation sequencing.

Another important data layer for prospective patients is digitalised histopathological images from histopathological samples collected during surgeries. A key aim of DECIDER is to produce at least four digitalised hematoxylin and eosin (H&E) stained histopathological images for each consented patient. The histopathological images are used in conjunction with sequencing data to identify drug resistance mechanisms as well as to improve diagnosis and predict treatment response. For a subset of the patients, tumour burden is measured with (18)F-fluorodeoxyglucose Positron Emission Tomography - Computed Tomography (FDG-PET/CT) imaging technology. The purpose of the FDG-PET/CT imaging is to accurately measure tumour burden before and after chemotherapy.

In addition to clinical, sequencing and imaging data from the prospective patient cohort, we will utilise a retrospective HGSOC patient cohort from the HUS-group, mice experiments, as well as data generated from organoids (established from prospective patient material) and commercial ovarian cancer cell lines.

The main purpose of the retrospective HGSOC cohort is to validate the results emerging from the prospective cohort in an independent validation cohort. The main measurement technologies to be used in the retrospective cohort are immunohistochemistry and RNA-in situ hybridisation staining that allow quantification of protein and gene expressions. However, some sequencing for individual genes relevant to drug resistance (e.g., *BRCA1/2*, *CCNE1*, etc.) may be conducted with the validation.

Organoid and mouse experiments are conducted to perform functional validation of the results as well as to test the efficacy of the suggested treatment options emerging from the project. The main measurement technology used in these experiments is imaging to quantify the tumour burden (mice experiments) and effect of the treatment (organoid experiments). When needed, the cells will be subjected to sequencing experiments to confirm gene mutation, copy-number, expression, or methylation status. Commercial cell lines are used to study the regulatory genome affecting HGSOC progression and drug resistance using single-cell RNA-sequencing and chromatin immunoprecipitation (ChIP) sequencing data.

2.2 Types and formats of data

2.2.1 Types and formats of research data collected in the project.

Clinical data from HGSOC patients

- Prospective clinical data from consented patients: patient personal information (name, social security number, municipality, age) and data on diagnoses, height and weight, surgical procedures, PET/CT imaging results, information on chemotherapy and other treatments, blood sample results, histopathological analyses such as IHC staining made in diagnostic routine, treatment outcome and survival. Clinical data are stored in a FileMaker Pro database managed by personnel in the TYKS. Pseudonymized clinical data exports for research use are made periodically from the clinical data database. Pseudonymized clinical data exports are shared with the DECIDER members who are authorised by the OPM (operational project manager) through eDuuni, which is a collaboration service environment for flexible and secure collaboration across organization and ecosystem boundaries provided by the government-owned company CSC — IT Center for Science. eDuuni is a service environment maintained by the Finnish state security regulation increased level (Vahti 2/2010), which is ensured with regular audits by CSC and external auditors. On this basis, eDuuni service environment can be used for material that is in protection level IV (Restricted) according to Finnish government decree 1109/2019.
- Retrospective clinical data from biobank: survival and recurrence times, stages, ages, routine longitudinal diagnostic laboratory, surgery, and treatment data. Data are stored in a FileMaker Pro database maintained by the personnel in HUS. The clinical data for the retrospective cohort are shared via eDuuni in a similar fashion as prospective clinical data.

Sequencing data

- Tissue, ascites, and plasma samples from consented patients are obtained during routine operations and sequenced. We will obtain whole-genome sequencing (WGS), RNA-sequencing, DNA methylation sequencing, circulating tumour DNA (ctDNA), shallow sequencing (plasma samples only) and exome-sequencing (plasma samples only) data. During the project, we may include other data layers pending technological advances in sequencing technologies.
  - WGS: FASTQ (raw read data). Downstream formats include BAM (mapped read data, processed read data as input for downstream analyses), VCF (variants), CSV/TSV (tab-separated tables for various types of data), and relational database management systems for combining various types of downstream data.
  - RNA-seq: FASTQ (raw read data). Downstream analysis results in BAM files, gene level and transcription level effective counts, and log2TPM expression data as csv format files and additionally relational database management systems for integration with genomic data.
- Sequencing data from cell lines and organoids: whole genome and exome DNA sequencing, DNA methylation, and RNA sequencing as above for tissues, genetic screening DNA sequencing as called genetic variants, ChIP-seq as FASTQ, BAM and CSV/TSV for variants and estimates from the data.
- Sequencing data from commercial ctDNA reference standards and commercial ctDNA reference standards spiked-in into control samples: FASTQ (raw read data). Downstream

formats include BAM (mapped read data, processed read data as input for downstream analyses), VCF (variants), CSV/TSV (tab-separated tables for various types of data).

Imaging data

Histopathological images from prospective and retrospective HGSOC patients

- H&E stained slides from prospective patients are obtained during routine operations from the consented patients. These slides will be scanned with pseudonymised sample codes in the Auria biobank and the digitalised H&E images with pseudonymised metadata are delivered with secured hard-disks to the relevant parties in the consortium. Retrospective samples from Helsinki Biobank are imaged at the biobank. Relevant histopathological slides are subjected to multiplexed immunohistochemistry (mIHC) or RNA-in situ hybridisation (RNA-ISH) experiments to validate the expression of a group of genes or proteins.
- RNA in situ hybridization images: Chromogenic and immunofluorescence RNA-ISH image (whole slide or tissue microarray) are in mrxs format files.
- Immunohistochemistry images: Chromogenic or immunofluorescence IHC images are in mrxs format files.
- Pseudonymised images of histopathological slides for the development of AI diagnostic tools at Aiforia are processed in their production environment hosted in Microsoft Azure (Western Europe data centres).

Radiological images from prospective HGSOC patients

- PET/CT and CT scans are obtained from consented patients as part of the preoperative workup and treatment monitoring.. These data are analysed by experienced radiologists in TYKS. The exact file format depends on the PET and CT scanner manufacturer, but common formats include DICOM, NIfTI, Interfile, ECAT, Analyze and NRRD. When CT DICOM images are converted to pseudonymized NIFTI or NRRD format on the TYKS server, those file formats no longer contain patient information in their header, but only information about image resolution and image reconstruction. Common image file formats (JPEG, TIFF, GIF, etc.) can be generated from PET and CT image files, but those are for viewing only as they no longer include quantitative information. The primary data from the images that are used in the project and saved in the patient clinical database are the results from the radiologist's statement. The actual image data is primarily stored in the hospital's PACS system, from where it is downloaded to Auria server and analysed. From these processed images, features like tumour volume and radiomics features can be generated with radiomics algorithms. These results are stored in the clinical database.

Imaging data from mice and organoids

- Images from mice and organoid experiments are saved as TIFF/JPEG images.

Measurement data from experiments and analyses

- Mice experiments are used in the validation of the results. Measurement data from mice include tumour growth measurements, i.e., tumour diameter as a function of time.
- Measured and collected data from organoid growth such as success of the organoids and performed validations are collected to eDuuni.

- Statistical summaries and other results from sequencing analyses: association effects and significance between genetic, epigenetic or transcriptional markers, and various biological features (e.g., drug resistance, disease stage, etc.).

The analysis subprojects (such as WGS data analysis, bulk RNA data analysis, histopathological image analyses, etc.) each have their own page within the project's internal wiki. These pages include information on the people responsible for the analyses, goals, data, methods, status and main results, at the minimum. An example of a data analysis workflow for short variant calling from WGS data is shown in Figure 1.

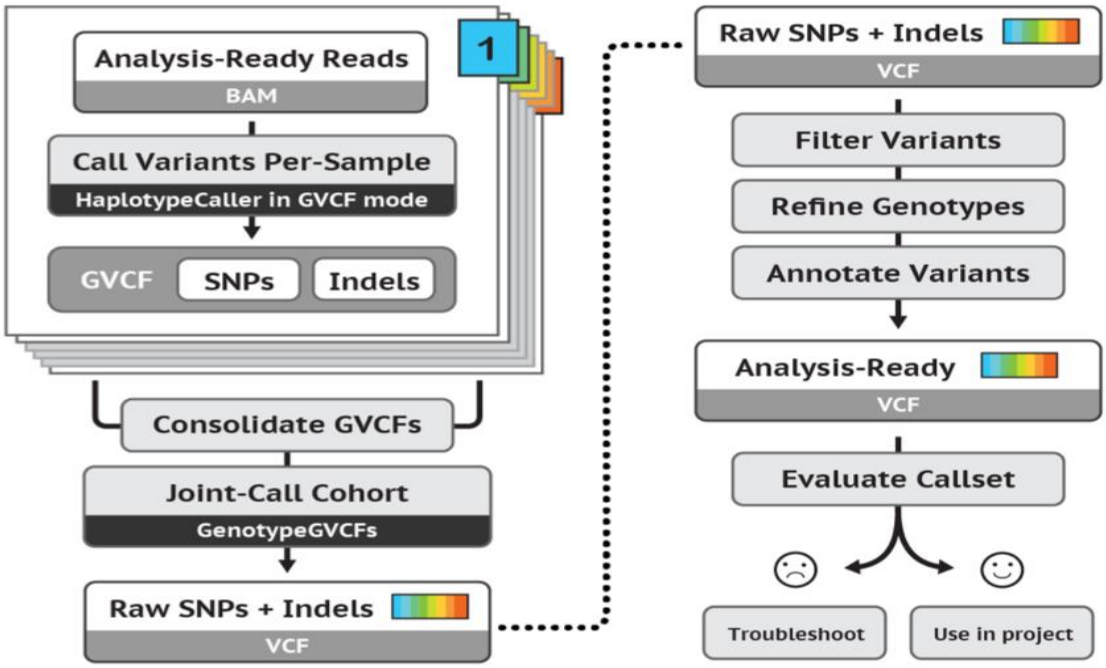

Figure 1. Workflow for calling germline short variants from whole genome sequencing data.

2.2.2 Data collected or generated for project management

All final signed project-related agreements, ethical permits and reports are archived as pdf documents (in addition to possible printed copies), and agreement templates, etc. as Word (.docx) documents.

2.3 Re-use of existing data

In addition to prospective and retrospective data utilised in DECIDER, we will take advantage of other large sequencing efforts that provide clinical and molecular data, such as The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), as well as other studies and scientific articles, for validation of the results emerging from DECIDER. TCGA and ICGC have their own Data Access Committee and we have been granted rights to use them. Access to other data repositories, such as BCCancer Canada and Sanger CGP, will be applied for when their use becomes relevant in DECIDER. The existing data have been mapped to various genome builds and we will use the genome assembly GRCh38 for all samples in DECIDER.

2.4 Origin of data

Clinical data

- Prospective clinical data: The data are collected from electronic hospital records in the Wellbeing Services County of Southwest Finland (formerly Hospital District of Southwest Finland), Satasairaala Central hospital, and Vaasa Central hospital, surgical operation notes and sample collection forms.
- Retrospective clinical data: The data related to the Helsinki Biobank samples are from the Biobank (provided for researchers in pseudonymized form).
- TCGA, ICGC and other large consortia provide clinical data, (typically diagnosis and survival, age, stage, etc.), which will be used in validation of the results from DECIDER.

Sequencing data

- Sequencing data from prospective patient samples: The samples are collected at the Wellbeing Services County of Southwest Finland/TYKS and sent for sequencing to an external sequencing provider (procurement of the service is performed periodically).
- Sequencing data from external repositories: Examples include but are not limited to the TCGA and the ICGC projects.
- Sequencing data from cell lines: The cell lines are procured from commercial sources and sequenced by various service providers or in-house.
- Sequencing data from organoid lines: The organoid lines are produced from the prospective patient tumours at Danish Cancer Society and sequenced by the company with whom the coordinator has a procurement contract with at the time.
- Sequencing data from commercial ctDNA reference standards and commercial ctDNA reference standards spiked-in into blood samples from healthy donors: reference samples are purchased from a commercial supplier, sequencing is performed in-house.

Imaging data

- Imaging data from PET/CT scans is produced by the Turku PET centre.
- Imaging data from histopathological slides of prospective samples is produced by Auria biobank, TYKS/Department of Pathology and University of Turku/ Histocore.
- Drug-sensitivity data from prospective samples will be obtained through automated screening and imaging using AI developed image algorithms by 2cureX.
- Imaging data from retrospective samples: Helsinki biobank slides are imaged at the biobank. RNA-ish and IHC images are expected to be generated using 3DHISTECH Pannoramic 250 FLASH II digital slide scanner at Genome Biology Unit core service at University of Helsinki.
- Imaging data from mice/organoids: automatic image analysis by Danish Cancer Society.

Documentation for AI tools

The AI tools are developed, and documentation provided by University of Helsinki, University of Modena and Reggio Emilia, Aiforia, Institut Pasteur, and Heidelberg University Hospital partners. The documentation will be available in the DECIDER wiki pages and upon publication

in GitHub or other openly accessible repositories. An example of the level of documentation is available in https://github.com/PrismLibrary/Prism.

Data collected or generated for project management

Agreements between partners and / or companies, where services are procured from, are produced by the partner's legal departments. Reports are produced by the partners responsible for the respective tasks / deliverables and reviewed by the coordinator before submission. Ethical permits are applied from the relevant authorities by the partners requiring them for the work.

2.5 Expected size of data

We expect to have data from approximately 350 HGSOC patients (prospective cohort). The below table presents an overview of the number of sample numbers per data category at the beginning of the project, as well as the set target numbers.

Table 1. Summary of main data levels, sources, and amounts collected and produced in the project.

Data category	Owner	# Beginning of the project	# Target	Information
HGSOC patients recruited to prospective cohort	TYKS	180	350	Approximately 60% are PDS patients and 40% NACT patients
Histopathological images (prospective cohort)	TYKS / UH	500	4,000	>10 digitalised images per patient from multiple sites and at different treatment stages
HGSOC patients in retrospective cohort	HUS/UH	900	1,500	FFPE tissue samples from patients in tissue microarray format. 3-6 sample-cores per patient from multiple sites.
Clinical data for prospective cohort	TYKS	180	350	Contains all routine longitudinal diagnostics and follow-up data. Additional data from surgeries and toxicities.
Clinical data for retrospective cohort	HUS/UH	700	1,500	The clinical data contains necessary information, such as survival times, stages, ages, routine longitudinal diagnostic laboratory, surgery and treatment data *etc*., for treatment prediction.

Data category	Owner	# Beginning of the project	# Target	Information
Tissue samples with whole-genome sequencing (WGS)	TYKS / UH	500	2,000	Samples from diagnosis, interval surgery & relapses (when possible)
Samples with RNA-seq	UH	500	2,000	Samples from diagnosis, interval surgery & relapses (when possible)
Samples with DNA methylation sequencing	UH	50	1,000	Samples from diagnosis, interval surgery & relapses (when possible)
In vitro / *ex vivo* samples with single-cell RNA-seq	KI	8	20	We aim primarily to use *ex vivo* samples, *in vitro* samples used as a backup plan
PET/CT functional imaging data (# of patients)	TYKS	30	60	Objective to have PET/CT at diagnosis and at interval surgery for the same patient.
CT imaging data (# of patients)	TYKS	25	350	Objective to measure tumour burden and radiomic parameters from clinically relevant time points
ctDNA sequencing data (# of patients)	TYKS / UH	30	300	Shallow-sequencing to see that there is enough material for exome-sequencing/large panel.
Molecular and clinical data from publications, TCGA, ICGC, *etc*.	—	1,300	1,800	We systematically collect HGSOC data from publications and repositories. In DECIDER such data is used for validation purposes.

Clinical data

- Prospective clinical data: The database consists of a multiple FileMaker Pro 19 database files, totalling approximately 470 MB in size. The expected number of patients, from whom data is collected, is estimated to be 350.
- Retrospective clinical data: The expected number of patients, from whom data is collected, is estimated to be 1,500.

Sequencing data

- Sequencing data from prospective patient samples:
  - WGS: 2,000 samples, approx. size 300TB.

  - RNA-seq: 2,000 samples, approx. size 50TB.
  - DNA methylation sequencing: 1,000 samples, approx. size 100TB (BAM files).
  - ctDNA sequencing data: 550 samples, approx. size of data 1 650 GB (FASTA files).
  - single-cell RNA-seq: 20 patients / cell culture samples, approx. size of data 300GB.
- Sequencing data from patient tumour organoids approx. size 10TB.
- Sequencing data from commercial cell lines:
  - Karolinska Institutet: 3 commercially available HGSOC cell lines will be used for generating scRNA-seq data after temporal and combinatorial drug treatments. Size estimate: 6TB.
    IRB Barcelona: 2-4 isogenic pairs of cell lines, 5-50 genotypes examined in each pair. (i) WGS data and (ii) genetic screening sequencing data. Size estimate: 10TB.
- Sequencing data from commercial ctDNA reference standards and commercial ctDNA reference standards spiked-in into control samples: 300 samples including serial dilutions, size estimate 30GB.

Imaging data

- Imaging data from PET/CT scans: Estimated for approx. 30 patients (in neoadjuvant treatment), imaged twice. Currently, only the results of the radiologist statement are saved for project use in the clinical database. If needed later, the images can be recovered from hospital/PET centre records.
- Imaging data from CT scans taken as part of clinical routine: estimated 350 patients. One CT scan contains multiple imaging series and totals approximately 1-3 GB. Image data is not necessarily stored as is in clinical database, as they can be retrieved from hospital PACS later on. Approximate size of the dataset is 500 to 1,000 GB. Image segmentation data is around 150 MB per study, meaning approximately 100 GB for 350 patients, if two segmentations are done per patient. The segmentation data is stored in Auria server.
- Imaging data from histopathological slides of prospective samples: 4,000 slides/images, approximate size of data 20TB.
- Imaging data from retrospective samples: data from 1,500 patients, 3-6 sample-cores per patient from multiple sites, approx. size 5TB.
- Imaging data from mice / organoids / tumoroids: estimate size of data 10TB.

Measurement data from experiments and analyses

- Mice experiments: Estimated size of data 2TB
- Statistical genomic association studies - statistical summaries of data. Estimated size of data <1 GB in total.

Data collected or generated for project management

- The documents are saved as Word and pdf files, approx. size 60 MB.

2.6 Data utility

We expect that the data based on full omics profiling of patients will be used by clinicians in clinical decision making. Based on similarities of drugs and drug targets, we attempt to predict statistically the synergy of drug combinations at various doses, which should be of interest for drug developing companies.

Clinical data

- Prospective and retrospective clinical data: Ethical permits and data protection regulations limit the amount of clinical data that can be shared outside the project, but the data that can be shared through publications and repositories in connection with e.g., the sequencing data are extremely useful for other ovarian cancer researchers.

Sequencing data

- Most sequencing data from prospective patient samples, including organoids, are also subject to ethical and data protection regulations but will be a valuable resource to other (ovarian) cancer researchers after they are archived to a controlled access data repository European Genome Phenome Archive (EGA).
- One of the features of DECIDER is to conduct a Dream Challenge competition in which anonymised data will be used. These data will be freely available and useful for method developers participating in the challenge.
- Sequencing data from retrospective patient samples will be stored to biobank, which allows the use of it for other researchers.
- Sequencing data from commercial cell lines: WGS data and the genetic screening data are anticipated to be made available as mutation calls (WGS) and read counts (screening), deposited in repositories such as FigShare or Dryad or similar. This should be useful to human genomics researchers and to cancer researchers.
- Sequencing data from commercial ctDNA reference standards and commercial ctDNA reference standards spiked-in into blood samples from healthy donors: the data are useful only for performance verification of assays developed within the project.

Imaging data

- Imaging data from PET/CT scans: The images themselves are not shared outside the project's consortium. Image features, such as tumour volume and locations of the masses, as well as radiomics parameters generated from routine CT scans are useful for mathematicians who develop models for tumour progression and can be used as part of DECIDER multiomics.
- Imaging data from histopathological slides of prospective samples: Histopathological images will be available from a constrained repository in which the Data Access Committee can ensure that the research questions are in line with the patient consent form. The consortium is looking into possible repository options as histopathological image repositories at the moment are scarce. An option is to utilise the federated repositories to be implemented in the H2020 INCISIVE or IMI BIGPICTURE projects in which University of Helsinki & HUS are also partners. Histopathological images will be useful for researchers and companies conducting digital pathology research and product development.
- Imaging data from retrospective samples will be stored in the biobank. Histopathological images will be useful for researchers conducting digital pathology research.
- Imaging data from mice, tumoroids, and organoids will be available in the publications and are useful for cancer researchers.

Data collected or generated for project management

Project reports that do not include any confidential data will be published through the EU CORDIS pages and can be useful e.g., for other project managers and to the general public interested in the project.

3. Findable, Accessible, Interoperable and Re-usable (FAIR) data

The DECIDER project is committed to making its research data findable, discoverable and identifiable. “Metadata” is structured information describing the characteristics of a resource, for example, the dates associated with a dataset. Metadata supports discovery, re-use and long-term preservation of resources. Metadata needs to vary across scientific fields but typically cover the following: i) descriptive metadata, such as title, abstract, author, and keywords; ii) administrative metadata which provide information that help manage a source, such as date of creation, file type and other technical information, like access rights; iii) Archive terms and access policies.

A metadata record consists of a set of predefined elements that define specific attributes of a resource. Each element can have one or more values; for example, a dataset may have multiple creators or more keywords may be added to a particular image to enable its finding. Documenting data enables other researchers to discover the data. Metadata about the nature of the files is also critical to the proper management of digital resources over time.

3.1. Making data findable, including provisions for metadata

Here we provide initial information regarding the application of FAIR principles to research data identified at this stage.

Clinical data from prospective cohort

Clinical data are coded with pseudonymised identifiers. Personnel in TYKS conduct periodic clinical data exports that ensure that the important clinical data features for researchers, such as survival times, are updated and usable. There are no explicit version numbers due to the nature of clinical data accumulating continuously.

Sequencing data

Sequencing data are coded with pseudonymized identifiers and are available via the EGA for researchers to which permission is granted given that their research plan is in line with patient consent, see section 3.2). The format of the data is <*patient code*>_<*time+tissue*>. As an

example, “H014_pOva” means that sequencing data are from a patient with the pseudonymised identifier H014, samples are taken in the primary treatment phase (diagnosis) from ovary tissue. Different sequencing platforms are also indicated as some samples have been sequenced with multiple sequencing protocols (for technical reasons, such as to reduce batch effect). For WGS, the FASTQ files have the following naming convention: *<sample_name>_<sequencing_platform>_seq<set_number>_<flowcell>_<lane>_<library>+<index>_<1/2 for read/mate>.fq.gz*. The pseudonymized ID in the sample name allows connecting the data to metadata (clinical data) within the project. For data protection reasons, in publications and for the archiving in the EGA the sample names are given using separate publication IDs. Sequencing platform, protocol, and data analysis metadata are provided in the EGA. Sequencing data originating from cell lines (WGS and genetic screening data) are anticipated to be annotated at least with the cell line name, genotype resulting from editing, screening conditions (e.g. drug concentration), and biological response (growth inhibition).

Imaging data

Histopathological image data from prospective patients have pseudonymized IDs followed by the tissue site with the same naming convention as sequencing data. This allows linking image data to sequencing data and clinical data. Retrospective patient image data will follow the naming convention in the biobank.

3.2. Making data openly accessible

All patient-related data are classified as sensitive and therefore cannot be shared freely. To comply with the Horizon 2020 pilot for open access to research data (ORD pilot) we need to balance openness and protection of scientific, and especially patient-sensitive information. The best solution we have identified is to share sensitive data via the European Genome-phenome Archive (EGA). The EGA is a service for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for the purposes of biomedical research projects. In the EGA, our sequencing data is organised in smaller datasets, and researcher can apply to the datasets via an online portal. Clinical data of the project is not shared via EGA. An internal Data Access Committee reviews applications from the researchers to judge whether their research questions are in line with the original patient consent form. Furthermore, researchers need to sign a Data Access Agreement to ensure that the sensitive data are not distributed further or accessed by persons without authorisation. Lawyers and the data protection team from the University of Helsinki support the Data Access Committee in questions reading additional agreements that might be needed for applicants with certain countries of origin.

To increase the usability of the sequencing data, we provide data sets that are cleaned so that they comply with EU legislation. For example, we provide summary data for WGS and RNA-seq as supplementary material to our scientific articles.

All our scientific methods are freely available with documentation in open repositories, such as GitHub, BitBucket, or the DECIDER website.

3.3. Making data interoperable

To facilitate exchange, we will use standardised formats as much as possible. For clinical data we will use ICD-10 and SNOMED (Systematized Nomenclature of Medicine, Clinical Terminology) coding. Mathematical models will be shared as SBML (Systems Biology Markup Language) format. For sequencing data, we follow the GDC (NCI Genomic Data Commons) standards. Clinically applicable genomic aberrations (mutations, copy-number events, gene fusions, etc.) are categorised according to the ESCAT (ESMO Scale for Clinical Actionability of molecular Targets) classification standard.

3.4. Increasing data re-use (through clarifying licences)

The open-source code produced in the project will be published under the 2-Clause BSD licence or similar and made available through GitHub or Zenodo.

Sequence data archived to EGA will be available to third parties also after the project ends (provided the intended use complies with the patient consent and approval is granted by the Data Access Committee).

4. Allocation of resources

The costs related to data management are mainly the salary costs of the personnel who prepare the data, e.g., sequencing data for archiving in EGA, and open access publication costs. Open access costs for publications and the salary of project personnel involved in data management are covered by the project grant. The currently selected repositories for depositing data are free of charge.

Each research group/company is responsible for handling and publishing their own data, however, sharing of data (e.g., clinical, sequencing and image data) within the consortium is the responsibility of the coordinator. Access to pseudonymised clinical data and to sequencing data from patient samples is currently managed by the operational project manager, Ann-Christin Ostwaldt (Helsinki University). Access to identifiable personal data is managed by TYKS/Johanna Hynninen. She is supported by coordinator Elina Valkonen (TYKS).

5. Data security

The database for prospective clinical data and samples is located on TYKS servers, where only authorised personnel have access to with a user-specific username and password and through 2-factor authenticated VPN tunnel. Pseudonymised exports of the data needed by the project researchers are kept in the eDuuni workspace (https://info.eduuni.fi/en/services/workspaces/),

which is an audited, secure collaboration environment, provided and maintained by CSC — IT Center for Science (https://www.csc.fi/en/home), which is owned by the Finnish state and higher education institutions. Access to the DECIDER eDuuni folders is granted by the project manager for those needing access to clinical data after their organization has completed a DTA (Data Transfer Agreement) or MTA (Material Transfer Agreement) with TYKS. In addition, each individual user of eDuuni has to sign a non-disclosure agreement (NDA), committing to the instructions on how to handle the data. The access log is checked by the project manager periodically.

Patient tissue samples and plasma are sequenced by commercial sequencing providers. The sequencing provider is regularly tendered, and procurement contracts are signed, ensuring that the service provider complies with GDPR and all EU and Finnish data security regulations.

All patient sequencing data are kept on organization servers behind firewalls, with access granted only to authorised personnel after they have signed the above-mentioned NDA. Sequencing data from patient materials are shared from the University of Helsinki to other DECIDER partners only after the organisations have completed a DTA. Transfer of sequencing data is done via encrypted hard disks or via short lived transfer servers. CSC — IT for Science also provides computing power and solutions for running large-scale analysis in a secure environment, used by the University of Helsinki.

Figure 2 below summarizes the process for granting access to clinical and sequencing data from patients.

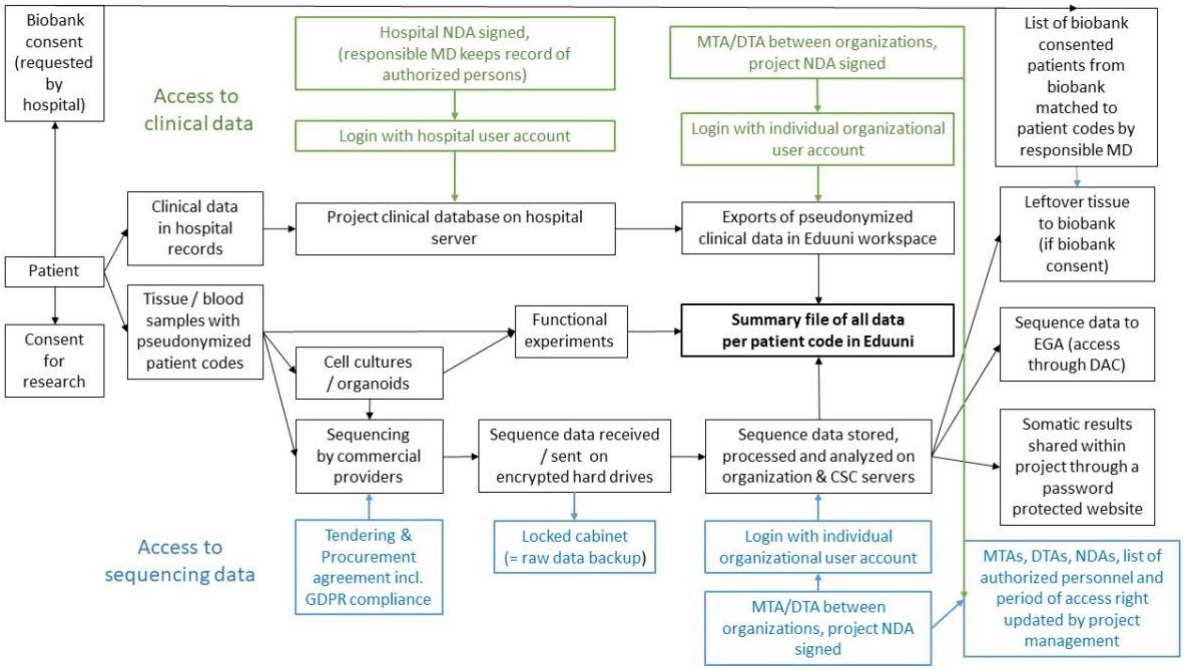

Figure 2. Prospectively collected patient data processing workflow. (Abbreviations: GDPR - General Data Protection Regulation, MTA - Material Transfer Agreement, DTA - Data Transfer Agreement, NDA - non-disclosure agreement, EGA - European Genome Phenome Archive, DAC - Data Access committee)

Further details of the organizational and technical measures to protect unauthorised access to sensitive data are described in Ethics Deliverable 13.4 (confidential).

5.1 Data storage and recovery

Clinical data are backed up daily, weekly and every 30 days. The number of respective backups stored is 24, 10, and 12.

Raw sequence data and histopathology images from prospectively collected patient samples arrive on hard disks, which are stored in a locked drawer. Data on the University of Helsinki servers are backed up and snapshots taken every second week. Sequencing data from commercial ctDNA reference standards and commercial ctDNA reference standards spiked-in into healthy control samples are backed up daily.

Sequencing data from prospectively collected patient samples will be archived in the EGA, specifically designed for long-term storage and sharing of this data type. Access to data is applied from the data access committee that reviews whether the applied use is in line with the original patient informed consent form. The data access committee consists of representatives of TYKS and University of Helsinki (owners of the data). Full clinical data related to the sequencing data cannot be shared due to personal data protection issues, but the publishable data will be available from linked publications.

5.2 Transfer of sensitive data

Pseudonymised clinical data is shared with the partners in the DECIDER project in a secure eDuuni workspace, provided by the CSC — IT Centre for Science, with access control and an audit log.

Patient sequencing data and imaging data from histopathological slides are transferred on encrypted hard drives as a tracked package with a courier (e.g., FedEx), with the decryption information only accessible in eDuuni for the sender and receiver.

Pseudonymised patient sequencing data can also be transferred to other DECIDER partners through the internet via short-lived transfer servers. The files are encrypted with GPG (GNU Privacy Guard), before being uploaded with SSH, to a cloud storage provider. The receiving partner downloads the data, after which the storage is deleted. Only the public SSH and GPG keys are transferred, leaving the decryption capability solely on the receivers' end.

Fully pseudonymised CT data in NIFTI or NRRD format can be transferred directly by authorised personnel from TYKS servers only to secure University of Helsinki servers (see section 5.1) for further processing.

6. Ethical aspects

Clinical and patient sequencing data fall under sensitive personal data and access to them must be restricted and monitored. These are covered in more detail by the (confidential) project ethics deliverables D13.1 (D52) H - Requirement No. 1, D13.2 (D53) H - Requirement No. 2, D13.4 (D55) POPD - Requirement No. 4, and D13.5 (D5) POPD - Requirement No. 5.

The patient information and consent forms for prospectively collected data include information on the possibility of the pseudonymized data being archived in repositories such as EGA.

The project involves four Small and medium-sized enterprises (SMEs) that aim to commercialize their results, therefore IPR issues must be considered before publishing any related data. The project has an innovation management panel to discuss these matters. All journal publications and abstracts of conference presentations/posters are also submitted for project review 30 days before publication, in order to review possible IPR issues.

7. Other issues

The DECIDER project involves the processing of sensitive data (health and genetic data) from data subjects that can be considered vulnerable (patients). Therefore, a Data Protection Impact Assessment (DPIA) for the project was drafted in January 2022 and revised by the Data Protection Officer of the University of Helsinki. The DPIA was updated to in January 2024.