DMP in the context of the FWO entitled Atlas of therapeutic targets in metastatic breast cancer.

ADMIN DETAILS

Project Name: DMP in the context of the FWO entitled Atlas of therapeutic targets in metastatic breast cancer.

Principal Investigator / Researcher: Christine Desmedt

Institution: KU Leuven

1. GENERAL INFORMATION

Name applicant

Christine Desmedt

FWO Project Number & Title

G082524N

Atlas of therapeutic targets in metastatic breast cancer.

Affiliation

KU Leuven

2. DATA DESCRIPTION

Will you generate/collect new data and/or make use of existing data?

It will be a combination of both. In UPTIDER (S64410, EC approved on the 30th of November 2020) breast cancer patients are participating to our post-mortem tissue donation program. We are retrospectively collecting clinical health data and some diagnostic archived bio material and upon the death of the patients, during the autopsy, we are collecting new biological samples from which new data will be derived.

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you

reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Please see table 1 below.

In general within the project:

- An electronic case report form (eCRF) data base, using the in-house REDCap available system, has been built by our laboratory in order to collect available patient's data from UZ Leuven data base (collected from the local patient's file) as well as data specifically retrieved for the study. The patient's data include demographic, pathological, clinical, treatment and outcome data. Data collected are in line with the GCP and GDPR requirements.
- Sample's data are collected within a lab management system called "lab collector". Data collected are in line with the UZ/KU Leuven biobank requirements.
- With regards to the prospective biological material collection, we will generate new data including histological and immunohistological measurements. In addition immune soluble markers will be assessed in the blood.
- In addition, shallow whole genome DNA sequencing (sWGS), whole exome sequencing (WES) and bulk RNA sequencing data will be generated.

Raw as well as processed data will be submitted to a public repository in the below described standard formats, to enable sharing and long-term validity of the data.

Data type	N° of cases	Data Source	Data content	Data Format	Volume (Go)
Clinical/ sample Data	30 patients (median of 32 samples per patient)	Patient file	demographic, pathological, clinical, treatment and outcome data	eCRF and lab collector	0.5
Histo(immunological) data	~1000	Generated	Markers score	Excel file and images	2.5
Soluble markers	30	Generated	Markers score	Text file	300
Bulk RNA seq	~1000	HiSeq 4000	Gene expression	.fastq, bam	2.5
SWGS	~1000	HiSeq 4000	Genomic profile	.fastq, bam	1
WES	~1000	HiSeq 4000	Genomic profile	.fastq, bam	10

Table1: Data Description

3. LEGAL AND ETHICAL ISSUES

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service

Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

Yes, we will use personal data. Data include demographic (age, gender), pathological, clinical, treatment and outcome data . At the time of UPTIDER project proposal we were not asked to fill in a PRET application (the project was set-up before the application was up and running and mandatory) however a GDPR form was filled in at the time.

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

The UPTIDER project was approved by the EC UZ/KU Leuven S64410.

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

There is potential tech transfer/valorization in the putative assays/biomarkers we measure or we aim to discover. We are working with the Leuven Research and Development department (LRD) and it is involved in all the Material Transfer Agreements and Data Transfer Agreements (MTA/DTAs) we have set-up in the context of this project.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

We are working with the LRD and LRD is involved in all the MTA/DTAs we have set-up in the context of this project. There is no specific restriction.

4. DOCUMENTATION AND METADATA

What documentation will be provided to enable reuse of the data collected/generated in this project?

All collected clinical data are stored either in an eCRF (REDCap). In REDCap, the documentation is automatically provided though the auto generated dictionary and code book where all the requested data are described. All collected data belonging to the sample collection are stored in an online lab management system called Lab Collector.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

A description tab is systematically added to the documentation in order to explain the header of the data collection sheets.

5. DATA STORAGE AND BACKUP DURING THE FWO PROJECT

Where will the data be stored?

The data will be stored on the KU Leuven servers. All systems but Lab Collector run on a secured and backed up server of KULeuven (managed by ICT of the Biomedical Sciences Group). These systems also provide a logging system so no data can ever be erased, making that everything will be traceable and stored long-term (well beyond the common 5-year requirement). Lab Collector data base is externally and professionally managed by the company "AgileBio".

How is backup of the data provided?

The hosting KUL server is automatically backed up using KUL services, multiple times per day. Concerning LabCollector, backups are automatically performed twice a day.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely. If no or insufficient storage or backup capacities are available then explain how this will be taken care of.

There is sufficient storage and back-up capacity on all KU Leuven servers:

- the "L-drive" is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp series storage systems, and a CTDB samba cluster in the front-end.
- the Staging and Archive on VSC are also sufficiently scalable (petabyte scale).

What are the expected costs for data storage and back up during the project? How will these costs be covered?

The total estimated cost of data storage during the project is 19200EUR. This estimation is based on the following costs:

- 8400 EUR for storage on the L-drive of 15000 Go at 700 EUR/5000Go/year over the total duration of the project (4 years)

- 10800 EUR for storage on the VSC cluster of 4500 Go at 0.6 EUR/Go/year for active data over the total duration of the project (4 years)

Budget for compute and data storage is budgeted for in ongoing projects (ERC and C1 granted proposals).

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Our eCRF, our lab management system (Lab Collector), and the network drive dedicated to the team (L-drive) are password access-protected by users, with person-based decision on rights to access and modify data. Moreover, this is the same for the VSC secured storage which is only accessible to VSC accounts, and specifically our volume will only be accessible to group members.

6. DATA PRESERVATION AFTER THE FWO PROJECT

Which data will be retained for the expected 5-year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).

The minimum preservation term of 5 years after the end of the project will be applied to all datasets. Biological samples obtained under research agreement will be kept according to the EC licenses and agreements. In effect, consent is obtained to store the samples for the specific research purposes stipulated in the informed consent. Putative remnant clinical/patient samples are stored in the UZ Leuven biobank.

Where will the data be archived (= stored for the longer term)?

As a general rule, datasets will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication or preprint deposition.

For all other datasets, long term storage will be ensured as follows:

- Large sequencing/omics data: will be stored on “L-drive”.
- Small digital files: files will be stored on the “L-drive”.
- Developed algorithms and software will be stored on L-drive, as well on public repositories such as framagit.com and codeocean.com.
- Clinical and sample data will be stored in our lab management tools (eCRF and Lab Collector).

Third-party software and algorithms that are used are referenced by their version numbers in our method section and are installed as modules on the VSC and/or containers (Docker, Singularity) on the VSC, to ensure reproducibility.

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

The total estimated cost of data storage during the 5 years after the end of the project is 19200 EUR. This estimation is based on the total given in Table 1 and the cost of storage on the L-Drive (700EUR/5000Go/year) and on the VSC cluster 0.6 EUR/Go/year. The cost will be covered by the laboratory budget.

7. DATA SHARING AND REUSE

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

Personal data will only be published after de-identification and identifiers will not be published.

Which data will be made available after the end of the project?

We are committed to publish research results to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible. Depending on their nature, some data may be made available prior to publication, either on an individual basis to interested researchers and/or potential new collaborators, or publicly via repositories (e.g. negative data). MTA and or DTA have been set-up in this sense and will be set-up if needed.

Where/how will the data be made available for reuse?

In an Open Access repository.

Upon publication, all anonymized patient details supporting a manuscript will be made publicly available as supplemental information.

Omics datasets will be deposited in open access repositories such the NCBI Gene Expression Omnibus (GEO) or The European Genome-phenome Archive (EGA). All the relevant algorithms, scripts and software code driving the project will be stored in a private online git repository of the laboratory. As soon as the manuscript is publicly available, the repository will be changed to a public repository.

When will the data be made available?

Upon publication.

Who will be able to access the data and under what conditions?

Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing.

Moreover, as mentioned above MTA and or DTA have been set-up and will be set-up if needed.

What are the expected costs for data sharing? How will the costs be covered?

It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible.

Data management costs will be covered by the laboratory budget.

8. RESPONSIBILITIES

Who will be responsible for data documentation & metadata?

Metadata will be documented by the senior post-doc, research manager, PhD students and technical staff at the time of data collection and analysis.

Who will be responsible for data storage & back up during the project?

The senior post-doc, research manager and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff.

Who will be responsible for ensuring data preservation and reuse?

The PI is responsible for data preservation and sharing, with support from the team, ICTS, gbiomed-IT staff, and UZ-IT staff.

Who bears the end responsibility for updating & implementing this DMP?

The PI is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.