Non-invasive Markers of an Early Response to Immune Checkpoint Blockade in Breast Cancer

A Data Management Plan created using DMPonline.be

Creator: Aurelie Mechels

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1S91723N

ID: 196571

Start date: 01-11-2022

End date: 31-10-2026

Project abstract:

Only a subset of breast cancer patients responds to immune checkpoint blockade (ICB). Predictive biomarkers are crucial to stratify patients most likely to derive clinical benefit from this expensive treatment. Existing biomarkers are often based on a single tumor biopsy, but the immunological profile of the tumor can change during therapy. Examining temporal changes requires serial tumor sampling, which is not standard practice. Our data on serial tumor biopsies showed that 30% of breast cancer patients develop intratumoral T cell expansion after ICB, which was used as a surrogate for response. Repeated blood sampling is minimally invasive, contains circulating immune cells, and several blood-based markers correlated with response to ICB in other cancer types. This project applies innovative single-cell profiling to pre- and on-treatment tumor and blood samples from early breast cancer patients with the aim to study the overlap between blood and tumor, and to identify a non-invasive predictive biomarker for response to ICB. This biomarker will also be validated in two independent breast cancer cohorts to assess its performance. This project will lead to more individualized therapy with maximum efficacy and minimal risks for patients. Moreover, it can reduce health care costs and thus help maintain a sustainable health care system.

Last modified: 11-04-2023

Non-invasive Markers of an Early Response to Immune Checkpoint Blockade in Breast Cancer Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

I will (re)use existing data and new data will be generated. Data will be processed and stored as FASTQ, TSV, HDF5, HTML, XLS, DOC, PPT, TXT, CSV, TIFF, PNG, or JPEG files on password-protected IT infrastructure. Biobanked patient samples will be labeled with sample IDs, and the key will remain with treating oncologists. Clinical data will be pseudonymized and linked to sample IDs before registration in RedCap. Papers will be published as open access via the KUL Lirias. If necessary, data will be reused by transfer through Belnet Filesender, and sequencing data can be uploaded on a public repository with appropriate access control. All data processing will be by the VIB-UZL/KUL SOPs, principles of GDPR 2016/679, and Belgian privacy law.

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

Biopsies for IHC and bulk RNAseq are stored at UZL Pathology. Other samples (blood/tumor tissue) are held at UZL Biobank. Dr. Bram Boeckx, a computer science expert with vast experience in data handling and use of the Flemish Super Computer (VSC) environment, handles sequencing data. Data is stored on servers with backup capacities managed by VIB-UZL/KUL. We use the ThinKing and Genius partitions of the VSC for handling raw sequencing data. Data are initially stored on the VSC staging and later on the archive for at least five years after the end of the research, both scalable up to petabyte-scale and under the Linux file system security.

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

I do not wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

There are no issues concerning research data indicated in the ethics questionnaire of this application form.

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

No other issues according to the data management are expected.

Non-invasive Markers of an Early Response to Immune Checkpoint Blockade in Breast Cancer DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

Non-invasive Markers of an Early Response to Immune Checkpoint Blockade in Breast Cancer GDPR

GDPR

Have you registered personal data processing activities for this project?

- Yes

Non-invasive Markers of an Early Response to Immune Checkpoint Blockade in Breast Cancer FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

The research and technical staff will generate, collect, process, analyze and store the data listed below, as detailed in the project description.

The following datasets will be generated:

In this project, we will analyze single-cell RNA (scRNA-seq), TCR (scTCR-seq), surface protein (CITE-seq), bulk TCR and RNA data . These data will be generated from tumor samples obtained from patients participating in clinical trials involving checkpoint immunotherapy.

Clinical samples listed below were already collected, and will be sequenced.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: - Generate new data - Reuse existing data	*Please choose from the following options:* - Digital - Physical	*Please choose from the following options:* - Observational - Experimental - Compiled/aggregated data - Simulation data - Software - Other - NA	*Please choose from the following options:* - .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … - NA	*Please choose from the following options:* - <100MB - <1GB - <100GB - <1TB - <5TB - <10TB - <50TB - >50TB - NA	
					Data will be processed and (temporarily) stored in the following formats: - Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format; Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS		

BioKey	Collection of 38 serial (pre- and on-treatment) tumor and blood samples ClinicalTrials.gov Identifier: NCT03197389	**Reuse existing data:** Tumor: scRNA-seq, scTCR-seq (76 samples, Bassez et al. (2021) Nat Med.) **Generate new data:** - Tumor: scBCR-seq (76 samples) - Blood: scRNA-seq, scTCR-seq, scBCR-seq (76 samples)	Both	Experimental	Excel (.xls/.xlsx);Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), PNG (.png), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; • Digital images in vector formats: scalable vector graphics (.svg), Adobe Illustrator (.ai), Adobe Portable Document Format (.pdf); • Next generation sequencing raw data: binary base call format (.bcl), .fastq Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig • Sequence alignment data: .bam • Structural variations data: .vcf, .bcf • Read/UMI count data: .tsv, .rds • Single-cell/nuclei suspensions and nucleic acid samples resulting from (single- cell) nucleic acid isolation, or sequence library preparations will be stored in labeled tubes or SBS plates in -20°C or -80°C freezers. Any leftover frozen tissue sections (each labelled with a study sample ID) will be stored in boxes in - 80°C freezers. We have electronic laboratory databases in .xls format that will keep the	< 5TB	76 tumor and 76 blood samples

					physical storage address of these samples.		
NeoImmunoBoost	Longitudinal tumor and blood samples from 52 patients collected at University Hospital Erlangen, Germany. ClinicalTrials.gov Identifier: NCT03289819	**Generate new data:** - Tumor: single nuclei RNA seq (snRNA-seq, n to be determined based on the amount of material left after bulk), bulk RNA, bulk TCR (112 samples) - Blood: scRNA-seq, scTCR-seq, CITE-seq (170	Both	Experimental	Data will be processed and (temporarily) stored in the following formats: - Text files: Plain text data (Unicode, .txt), MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format; Quantitative tabular data: comma-separated value files (.csv), tab-delimited file (.tab), delimited text (.txt), MS Excel (.xls/.xlsx); - Digital images in microscopy file formats: Nikon format (.nd2), Zeiss format (.czi), Leica format (.lif), Olympus format (.oib) - Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), PNG (.png), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif; - Digital images in vector formats: scalable vector graphics (.svg), Adobe Illustrator (.ai), Adobe Portable Document Format (.pdf); Next generation sequencing raw data: binary base call format (.bcl), .fastq Coverage data: .bed, .bg, .bedGraph, .bw, .bigwig	< 5TB	All blood and tumor samples are stored.

		samples)			• Sequence alignment data: .bam • Structural variations data: .vcf, .bcf • Read/UMI count data: .tsv, .rds • Spatial imaging count data: .txt, .csv, .mat • Single-cell/nuclei suspensions and nucleic acid samples resulting from (single- cell) nucleic acid isolation, or sequence library preparations will be stored in labeled tubes or SBS plates in -20°C or -80°C freezers. Any leftover frozen tissue sections (each labelled with a study sample ID) will be stored in boxes in - 80°C freezers. We have electronic laboratory databases in .xls format that will keep the physical storage address of these samples.		

Estimated volume raw data (.fastq file and .bam file) per work package that will be stored for long-term:

WP1: scRNA-seq, scTCR-seq, scBCR-seq of paired pre- and on-treatment tumor and blood samples from 38 patients. Estimated volume raw data (.fastq file) ~ 30 GB per sample * 4 (4 samples per patient) * 38 (38 patients) = 4560 GB ~ 4.6 TB

WP3:

- scRNA-seq, scTCR-seq, CITE-seq of a total of 170 longitudinal blood samples. Estimated volume raw data (.fastq file) ~ 30 GB per sample * 170 samples = 5.1 TB
- Bulk RNA and TCR on 112 longitudinal tumor samples. Estimated volume raw data (.fastq file) ~ 3 GB per sample * 112 samples = 336 GB
- snRNA-seq on leftover tumor material ~ 10 GB per sample * 54 (we will aim for pre- and on-treatment from 27 patients who received the initial boost, see project proposal) = 540 GB

Total estimated volume raw data = 10.5 TB

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

Biokey tumor scRNA-seq and scTCR-seq https://www.nature.com/articles/s41591-021-01323-8, https://doi.org/10.1038/s41591-021-01323-8
ClinicalTrials.gov Identifier: NCT03197389

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

Biokey:

Biokey is a monocentric clinical trial, with UZ leuven as patient recruitment site. Ethical approval for use of Biokey samples and patient data is already in place, and was approved by the Ethics Committee Research UZ/KU Leuven. Appropriate informed consent was obtained from all patients included in the trial.

EudraCT Number: 2016-004859-77
ClinicalTrials.gov Number: NCT03197389
Sponsor: UZ Leuven
Collaborator: Merck
S-number: S60100_003
Approval date: 03/08/2018

NeoImmunoBoost:

NeoImmunoBoost is a phase II, multi-center, neoadjuvant clinical trial conducted primarily at university hospital Erlangen (Germany) and is supervised by prof. dr. Peter Fasching.

EudraCT Number: 2016-003102-14
Antrag Number: 130_17 Az
ClinicalTrials.gov Number: NCT03289819
Sponsor: Institut für Frauengesundheit GmbH, Universitätsstrasse 21-23, 91054 Erlangen
Approval date: 06/12/2017

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

We will generate (single-cell) transcriptome sequencing data from from serial tumour samples from breast cancer patients participating to 2 different clinical trials involving checkpoint immunotherapy (Biokey ClinicalTrials.gov Identifier NCT03197389; NeoImmunoBoost: ClinicalTrials.gov Identifier: NCT03289819). The patient samples that we will receive for sequencing are coded with a study sample ID (the coding key remains with the oncologists, UZ Leuven/University Hospital Erlangen). This code does not carry any personal identifiers, keeping the identity of the study participant private and confidential. Access to the coding key is necessary to link any data or biological samples back to a subject identifier. In addition, when the samples are processed in our lab, they also receive a DILA ID (double-coded) and this DILA ID is further used in the downstream analyses. We will also receive pseudonymized patient data linked to the sample IDs. The clinical patient data will include the following information:

For Biokey: patient number, sample ID, sample type (pre/on-treatment), biopsy type, gender, inclusion date, age at diagnosis, cohort (upfront surgery or pre-treated with chemo), histological type, tumour grade, Ki67 at core biopsy, Ki67 at resection biopsy, pTNM, cTNM, HER2 status, TILs at core biopsy, TILs at resection biopsy, pre/post-menopausal state, BRCA status, breast cancer subtype, histological type, previous medical history, concomitant medication, BMI.

For NeoImmunoBoost: patient number, sample ID, sample type (Baseline, after initial boost, before Epirubicin/Cyclophosphamide, at surgery), biopsy type, gender, cohort (with or without initial boost), age at diagnosis, pTNM, cTNM, ER status, PR status, response, combined positive score for PD-L1, TILs at core biopsy, TILs at resection biopsy

The sequencing data that will be generated within this project will be correlated with the pseudonymized clinical data. The sequencing data and the associated pseudonymized patient data are defined as sensitive personal data and will only be processed in accordance with the institutional SOPs, the principles of the General Data Protection Regulation (GDPR) 2016/679 and the Belgian privacy law. These procedures include procedures for pseudonymization, data storage and data protection. The data subject will give consent for the processing of those personal data for the purpose of the current proposal. The data will be processed and stored on the institutional password protected IT infrastructure which is protected by a genuine user authentication system relying on username and password. Access to the data as well as the access level will be limited on a project need and individual basis. Only the researchers working on the project has access to these data. Due to the sample labeling as protective measure, the researchers are not able to decipher the identity of the donor.

All data that will be collected and the strategy to guarantee the privacy of the study participants are specified in the research protocol approved by the ethical committee.

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- Yes

We do not exclude that the proposed work could result in research data with potential for tech transfer and valorization. Both VIB and KU Leuven have a policy to actively monitor research data for such potential. If there is substantial potential, the invention will be thoroughly assessed, and in a number of cases the invention will be IP protected (mostly patent protection or copyright protection). As such the IP protection does not withhold the research data from being made public. In the case a decision is taken to file a patent application it will be planned so that publications need not be delayed.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

The work described in the current proposal is included as the translational research part of the 2 clinical trials (Biokey and NeoImmunoBoost). In the current proposal, we will use patient samples and data collected from patients participating in these trials at UZ Leuven and University Hospital Erlangen. Therefore, there is a Material and Data Transfer Agreement between the legal entities of KU Leuven (> UZ Leuven) and VIB; and between University Hospital Erlangen and VIB.

Moreover, non-commercial research agreements between the collaborator (Merck) of the Biokey trial is already in place. These 3rd party agreements will not restrict dissemination or exploitation of the data we will generate within this project. In addition, existing agreements between VIB and KU Leuven do not restrict publication of data.

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- Yes

Parties have expressly agreed that any and all data as collected and prepared in the context of this study shall be the joint property of UZ Leuven / KU Leuven and VIB.

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

- Documentation will consist of notes in the (electronic) laboratory notebook, the Grand Challenge database and metadata sheets, all to ensure that data files, lab samples, and experimental notes remain properly linked to the same study sample ID. The notes will describe the tumour samples used, the experimental setup, the protocols used, the physical location of the samples and the (sequencing) data generated. Samples will be stored in boxes labelled per clinical trial. Metadata sheets are maintained with the connection between lab samples (and their physical location), DILA IDs, study sample and patient IDs, the specific names of the respective datasets and a link to the sequencing runs (and their specific computer and physical location). A Grand Challenge database has been specifically created to enable visualization, organization and documentation of all samples processed in this project. This database contains clinical data, information about the specific clinical trials and contact persons, the tumour types, the patients included, the fresh and frozen tissue samples, the fresh and frozen sequencing libraries and other tissue available. Specific filters can be applied to verify and download only a subset of the data.
- All samples and corresponding files will be named following a standard procedure, so that all the name of all files in a given dataset will be in the same format. All changes in the files will be recorded, and deviations from the standard format will be added by _remark (ex. _frozen).
- Research methods and practices (SOPs) will be fully documented. When the wet lab techniques, scripts, algorithms and software tools are finalized, they will be additionally described in manuscripts and/or on GitHub.
- Raw sequencing data (.fastq, .bcl files and read count data matrices; each named with their DILA ID) will be stored on the Vlaamse Super computer (VSC), ordered in folders per sequencing-run, including an .xlsx file with the sample sheet information containing the DILA IDs sequenced in that run and the sequencing run information per DILA ID (Illumina sequencer, lane and index information). The name of the folder will contain the date of the sequencing run and the Illumina sequencer used. When data are published, raw and processed sequencing data will be uploaded on a public repository (e.g. EGA) with appropriate access control if required, to enable sharing and long-term validity of the data. Any data shared will only be released prior to a Data Transfer Agreement that will have to include the necessary conditions to guarantee protection of personal data (according to European GDPR law). Double/triple-coded read count data matrix (linked to double/triple-coded human data) will be available

on our website (https://lambrechtslab.sites.vib.be/en/data-access). Raw imaging data (microscopy files, imaging protocol details, count matrices) will be stored on password-protected and backed up VIB-KU Leuven IT infrastructure in folders per imaging run. The name of the folder will contain the date of the imaging run and the microscope and imaging protocol used.

All data will be processed and (temporarily) stored on secured, password-protected and backed up servers of VIB-KU Leuven (managed by ICT of the Biomedical Sciences Group) or on the VSC.

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Sequencing data types require specific metadata when submitted to public repositories such as EGA, ArrayExpress, GEO or ENA. Imaging data also requires specific metadata when submitted to public repositories such as OMERO or SpatialDB. Data documentation will be tailored to their ultimate deposition in public repositories, with spreadsheet headers corresponding to fields required by these public repositories. Technical and analytical methods used to generate the data will be documented in sufficient detail to allow for independent reproduction. These will include analysis package version numbers, analysis kit, disease status, treatment type and duration, organism, genome build.... For single-cell experiments, each droplet (cell) barcode will also be retained alongside the associated single-cell quality metrics. When depositing data in a repository, the final dataset will be accompanied by this information in the file format that the repository provides. This will allow the data to be understood by other members of the laboratory and add context to the dataset for future reuse.

3. Data storage & back-up during the research project

Where will the data be stored?

All electronical data collected and generated during the project will be processed and (temporarily) stored on secured, password-protected and backed up servers of VIB-KU Leuven (managed by ICT of the Biomedical Sciences Group).
The sequencing data generated during the project will either be stored on VIB-KU Leuven servers or on the Flemish Supercomputer Centre (VSC), initially in the staging and archive area, and later only in the archive area (archive is mirrored).
Raw and processed data will be submitted to a public repository (e.g. EGA) with appropriate access control if required, to enable sharing and long-term validity of the data. Double/triple-coded read count data matrix (linked to double/triple-coded human data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data- access).
All patient samples and their derivatives will be stored in labeled tubes or SBS plates in -20°C or -80°C freezers purchased by our own funding. The samples will be registered and handled according to the UZ Leuven Biobank guidelines, in compliance with the Belgian law on human body material (dd 19-12-2008).

How will the data be backed up?

KU Leuven drives are automatically (daily) backed up using KU Leuven services according to the following scheme:

- Data stored on the “L-drive” is backed up daily using snapshot technology, where all incremental changes in respect of the previous version are kept online; the last 14 backups are kept.
- Data stored on the “J-drive” is backed up hourly, daily (every day at midnight) and weekly (at midnight between Saturday and Sunday); in each case the last 6 backups are kept.
- Data stored on the digital vault is backed up using snapshot technology, where all incremental changes in respect of the previous version are kept online. As standard, 10% of the requested storage is reserved for backups using the following backup regime: an hourly backup (at 8 a.m., 12 p.m., 4 p.m. and 8 p.m.), the last 6 of which are kept; a daily backup (every day) at midnight, the last 6 of which are kept; and a weekly backup (every week) at midnight between Saturday and Sunday, the last 2 of which are kept.
- Incremental backups are done daily from one 20 TB QNAP NAS to a second 20 TB QNAP NAS.

All sequencing data stored on the Flemish Supercomputer Centre (VSC) will be transferred on a regular basis to the archive area which is backed up.
Data is stored on EGA/our website for the purpose of data sharing.

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.

- Yes

There is sufficient storage and back-up capacity on all VIB-KU Leuven servers:
- the “L-drive” is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp e-series storage systems, and a CTDB samba cluster in the front-end.
- the “J-drive” is based on a cluster of NetApp FAS8040 controllers with an Ontap 9.1P9 operating system.
- the Staging and Archive on VSC are also sufficiently scalable (petabyte scale).

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

Since we are working with personal data (pseudonymized genome data and associated pseudonymized patient data), the data will be processed and stored on the VIB-KU Leuven IT infrastructure which is protected by a genuine user authentication system relying on username and password: Both the “L-drive” and “J-drive” servers are accessible only by laboratory members, and are mirrored in the second ICTS datacenter for business continuity and disaster recovery so that a copy of the data can be recovered within an hour. Access to the digital vault is possible only through using a KU Leuven user-id and password, and user rights only grant access to the data in their own vault. Sensitive data transfer will be performed according to the best practices for “Copying data to the secure environment” defined by KU Leuven. The operating system of the vault is maintained on a monthly basis, including the application of upgrades and security patches. The server in the vault is managed by ICTS, and only ICTS personnel (bound by the ICT code of conduct for staff) have administrator/root rights. A security service monitors the technical installations continuously, even outside working hours.

Access to the data as well as the access level will be limited on a project need and individual basis. Only the researchers working on the project has access to these data. Due to the sample labeling as protective measure, the researchers are not able to decipher the identity of the donor.

No personal data will be stored on the VSC nor local drives, except for the nucleic acid sequences. The coding key to patient information of linked pseudonymized data will be kept with the oncologists of UZ Leuven.

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

The total estimated cost of data storage during the project is ~2500 EUR. This estimation is based on the following costs:

- €868,9/5 TB/Year for the “L- drive”
- €519/TB/Year for the “J-drive”
- The cost of VSC archive is €70/TB/Year and staging €130/TB/Year.
- Electricity costs for the -20°C/-80°C freezers present in the labs are included in general lab costs.
- Data backup costs are included in general lab costs.

We expect costs to drop slightly during the coming four years. The costs for data storage will be covered by our own funding.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

The work described in the current proposal is included as the translational research part of two different clinical trials involving checkpoint immunotherapy (Biokey ClinicalTrials.gov Identifier: NCT03197389 and NeoImmunoBoost ClinicalTrials.gov Identifier: NCT03289819) at UZ Leuven and University Hospital Erlangen, respectively. The datasets, collected in the context of clinical research, which fall under the scope of the Belgian Law of 7 May 2004, will be archived for 25 years, in agreement with the European Regulation 536/2014 on clinical trials of medicinal products for human use. Subsequently, they may be kept for an additional period of time, for the described scientific purposes of the clinical trials or for any legal reason (change of obligations with regard to storage, for example).

Where will these data be archived (stored and curated for the long-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) data will be stored on VSC archive
- Small digital files will be stored on the “L-drive”.
- Developed algorithms and software will be stored on VSC archive and/or L- drive, as well on public repositories such as Github.com

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

The total estimated cost of data storage during the 25 years after the end of the project is ~ €18.375. This estimation is based on 10.5 TB in total, at 70EUR/Tb/year. The storage after the project is much smaller because during the project a large working space is needed, and post-publication data will be made accessible via open access platforms. The costs for this data preservation will be paid by our own funding. Electricity costs for the freezers present in the labs are included in general lab costs.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)
- Other, please specify:

The PI in the present project is 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 at the latest at the time of publication (or preprint deposition) via the required link in the publication or upon reasonable request and after an embargo period after publication.

- Double/triple-coded raw and processed sequencing data (linked to double- coded patient data) will be submitted to a public repository (e.g. EGA) with appropriate access control. Relevant imaging and spatial analysis data (double/triple coded) will be deposited in specialized open access repositories with appropriate access control, e.g. OMERO or SpatialDB. Accession to these data will be made available to any individuals making a specific request and this request will be handled by the institutional data access committee (DAC). Any data shared will only be released prior to a Data Transfer Agreement that will have to include the necessary conditions to guarantee protection of personal data (according to European GDPR law). The double/triple-coded read count data matrix (linked to double/triple-coded patient data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data-access). Note: Personal data will be double/triple coded and no reference to subject name will be made.
- Scripts, algorithms and software tools will be described in manuscripts and/or on GitHub.
- The results will be published as BioRxiv preprints and as Open Access in peer reviewed journal.

If access is restricted, please specify 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. As detailed above, metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication or an ODC Public Domain Dedication and Licence, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. A CC-BY license will be opted for when possible. For data shared directly by the PI (and approval of the 3rd party if necessary), a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

For KU Leuven data submitted to the EBI European Genome-phenome Archive (EGA), which operates under controlled access, the data access/submission requests will be received by the Genomics Data Access Committee (DAC) of KU Leuven (https://homes.esat.kuleuven.be/~bioiuser/dac/) and processed in consultation with the PIs produced data. The DAC will provide general guidance in terms of policies and will be referred to in handling controversial cases.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Privacy aspects
- Yes, Ethical aspects
- Yes, Intellectual Property Rights

In general, personal data will only be published after de-identification and identifiers will not be published. If despite all efforts it is not possible to protect the identities of subjects even after removing all identifiers, personal data will not be made public.

In order to respect the patient’s privacy, tumour samples will only be available to the research and technical staff involved in the project, not to other groups, studies or purpose, unless ethical approval is granted.

We aim at communicating our results in top journals that require full disclosure of all included data, or restricted access through a repository with appropriate access control (e.g. EGA). Additional material or information could be shared upon simple request following publication, unless we identify valuable IP, in which case we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use.

The work described in the current proposal is included as the translational research part of the different clinical trials. In the current proposal, we will use tumour samples collected from patients participating in the Biokey trial at UZ Leuven. Therefore, we have a Material and Data Transfer Agreement (MTA) between the legal entities of KU Leuven (> UZ Leuven) and VIB; and between University Hospital Erlangen and VIB.

Moreover, non-commercial research agreements between the collaborator (Merck) of the Biokey trial is already in place. These 3rd party agreements will not restrict dissemination or exploitation of the data we will generate within this project.

The permission to share encoded data / samples is obtained in the informed consents which will be signed by the study participants before being included in the trial.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

Whenever possible, datasets and appropriate metadata will be made publicly available through repositories that support FAIR data sharing. Personal data will be double coded and no reference to subject name will be made. Sharing policies for specific research outputs are detailed below:

- Double/triple-coded raw sequencing data (linked to double-coded patient data) will be deposited in open access repositories with restricted access control such as the EBI European Genome-phenome Archive (EGA). The EGA is a repository for personally identifiable genetic and phenotypic data. Sequencing data at EGA will only be available upon reasonable request via our institutional data access committee and if necessary a material transfer agreement will be concluded with the beneficiaries in order to describe the types of reuse that are permitted. The double/triple-coded read count data matrix (linked to double/triple-coded patient data) will be available on our website (https://lambrechtslab.sites.vib.be/en/data- access).
- Double/triple-coded patient data: Upon publication, all double/triple-coded patient details supporting a manuscript will be made publicly available as supplemental information.
- Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents (raw data) deposited in the E-Notebook are accessible to the PI and the research staff and will be made available upon request.
- Manuscripts: All scientific publications will be shared openly. Manuscripts submitted for publication will be deposited in a pre-print server such as bioRxiv. At the time of publication, research results will be summarized on the PI’s website (https://lambrechtslab.sites.vib.be/) and post-print pdf versions of publications will be made available there if allowed by copyright agreements, possibly after an embargo as determined by the publisher. Before the end of the embargo or in cases where sharing the post-print is not allowed due to copyright agreements, a pre-print version of the manuscript will be made available. (Pre-print) publications will also be automatically added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID.
- Algorithms, scripts and software: All the relevant algorithms, scripts and software toosls driving the project will be described in manuscripts and/or on GitHub (https://github.com) and/or on our interactive webserver (http://blueprint.lambrechtslab.org).
- Extra data that do not support publication will be either deposited in an open access repository or made available upon request by email. Data will be reused by transfer via Belnet Filesender or secure copy.

When will the data be made available?

As a general rule all research outputs will be made openly accessible at the latest at the time of publication (or preprint deposition). No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed.

Which data usage licenses are you going to provide? If none, please explain why.

As detailed above, metadata will contain sufficient information to support data interpretation and reuse, and will be conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal

(CC0 1.0) Public Domain Dedication or an ODC Public Domain Dedication and License, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. A CC-BY license will be opted for when possible. For data shared directly by the PIs (and approval of the 3rdparty if necessary), a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

cfr. supra

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

It is the intention to minimize data management costs by implementing standard operating procedures (SOPs) 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 own funding.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

(Meta)data will be documented by the research and technical staff at the time of data collection and analysis, by taking careful notes in the (electronic) notebook that refer to specific datasets and by maintaining metadata sheets that preserve the connection between lab samples, sample and patient IDs, the specific names of the respective datasets and a link to the sequencing runs.

Who will manage data storage and backup during the research project?

The research and technical staff will ensure data storage and back up, with support from ICTS, gbiomed-IT staff, and UZ-IT staff. The project coordinator will regularly verify these protocols are followed.

Who will manage data preservation and sharing?

The PIs are in the end responsible for data preservation and sharing, in cooperation with the project coordinator, with support from ICTS, HPC, gbiomed-IT staff, and UZ-IT staff.

Who will update and implement this DMP?

The PIs bear the end responsibility of updating & implementing this DMP, supported by the project coordinator.