FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. **General Project Information**	
Name Grant Holder & ORCID	**Elena Donders ; 0000-0002-8466-7755**
Contributor name(s) (+ ORCID) & roles	**Els Wauters ;** 0000-0002-0115-0030 ; PI of project Pierre Van Mol ; 0000-0003-1314-9413 ; PhD student collaborating on project
Project number[1] & title	NCT04807114, S63531
Funder(s) GrantID[2]	ELR-FOREO1-O2010 (Fonds Respiratoire Oncologie); Postdoctoraal mandaat Stichting tegen Kanker Els Wauters (2017, 2022)
Affiliation(s)	☒ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible:

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

Please provide a short project description	The next wave of cancer immunotherapy involves combination regimens. Durable response rates are expected to rise for some, but this will come at a cost of unnecessary toxicity for others. In first-line non-small cell lung cancer (NSCLC), clinicians now have to decide whether or not to add CTLA-4 blockade to the current standard of anti-PD-1 and chemotherapy. However, scientific guidance to weigh risks and benefits of dual checkpoint blockade in a single patient is lacking, hampering a rational choice between different treatment options. By leveraging single-cell sequencing techniques in all relevant immune compartments, both before and during treatment, I will examine immune cell dynamics in advanced NSCLC patients treated with PD-1 blockade and chemotherapy with or without anti-CTLA-4. Particularly, I will assess primary tumor, draining lymph nodes as well as blood. This analysis will not only yield hypotheses on the immunological mechanisms underlying efficacy of available regimens, but also reveal predictive biomarkers. Via spatial transcriptomics, I will further investigate interactions within specific (predictive) immune neighborhoods. Finally, to assess clinical applicability, I will validate identified predictive immune signatures on FFPE samples from an independent NSCLC cohort using bulk RNA-sequencing and immunohistochemistry. By guiding rational selection of immunotherapy regimens, these findings can impact both clinical practice and future clinical trial design.

2. 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[4].

				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
ICB			☒ Digital ☒ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other ☐ NA	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☐ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: cfr. below ☐ NA	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☒ < 50 TB ☐ > 50 TB ☐ NA	Cfr. below
→ Generate new data – ICB (WP 1&2) The data listed below will be generated, processed, analysed and stored, as detailed in the research project.							

[4] Add rows for each dataset you want to describe.

A) Genomic, transcriptomic and proteomic *raw data*

Genomic, transcriptomic and proteomic data from patient samples will be generated locally using Illumina HiSeq4000 and NovaSeq6000 machines, producing files in .fastq or .bcl format.

We aim to generate data from tumour and blood samples from 60 patients (30 patients 2019-2021; 30 patients 2022-2024) treated with systemic therapy for advanced lung carcinoma at UZ Leuven. Each patient sample is labelled with a study sampleID, keeping the identity of the study participant private and confidential. Pseudonymized patient data (including basic demographics and clinical response/outcome data) linked to the corresponding sampleID will be recorded in .xls/.xlsx format.

Sequencing data generated in this project will be correlated with clinical response/outcome data. 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 raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif;
- Digital images in vector formats: scalable vector graphics (.svg), Adobe Illustrator (.ai);
- Next generation sequencing raw data: binary base call format (.bcl), .fastq
- Sequence alignment data: .bam
- Structural variations data: .vcf, .bcf
- Read/UMI count data: .tsv, .rds

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

WP 1: single-cell data

- Single-cell (sc)RNA-seq and scTCR-seq of pre-treatment tumour biopsies from 60 patients. Estimated volume raw data (.fastq file): ~30 GB (~10 GB scRNA-seq + ~20 GB scTCR-seq) per sample * 60 = 1.8 TB
- Single-cell (sc)RNA-seq and scTCR-seq of on-treatment tumour biopsies from +-10 patients. Estimated volume raw data (.fastq file): ~30 GB (~10 GB scRNA-seq + ~20 GB scTCR-seq) per sample * 10 = 0,3 TB
- Single-cell (sc)RNA-seq and scTCR-seq of serial PBMCs (2 timepoints per patient, 60 patients). Estimated volume raw data (.fastq file): ~30 GB (~10 GB scRNA-seq + ~20 GB scTCR-seq) per sample * 120 = 3,6 TB
- Whole-exome sequencing of serial plasma samples (2 timepoints per patient, 60 patients) and corresponding germline DNA. Estimated volume [.fastq file (~10 GB) + .bam file (~10 GB)]: ~20 GB per sample * 120) = 3,6 TB

Total estimated volume raw data = **9,3 TB**

WP 2: spatial data

- Spatial transcriptomics using Resolve Biosciences smFISH technology on pre-treatment tumour biopsies from 5 patients. Estimated volume raw data (count matrix delivered Resolve: **~13GB** per slide (4-8 samples)

B) Biological samples

Biological samples (including full blood, PBMCs, plasma, tumour tissue), single-cell suspensions and sequencing library preparations will be stored in labelled tubes or SBS plates in -20°C or -80°C freezers. Electronic laboratory databases in .xls format are used to keep track of the of these samples and their link to the original study sampleID. All biological samples are then registered and stored according to the guidelines of the UZ/KU Leuven Biobank.

Regarding volume we estimate:

- Tumor/LN biopsy pre-treatment: 1 piece x 60 patients = 60 pieces
- Tumor/LN biopsy on-treatment: 1 piece x 10 patients = 10 pieces
- PBMCs/full blood/plasma: 2x10mL x 60 patients x 2 time points = 2400mL

		☒ Reuse existing data					

→ **Reuse existing data**

As part of **WP 3** of this project we aim to use existing, publicly available datasets for validation experiments:

- The Cancer Genome Atlas Research Network (TCGA)
  - Wheeler *et al*. Cell. 2017.
  - Available in the GDC Portal : https://gdc.cancer.gov/access-data
  - Data type : .fastq file
  - Size: variable size, up to 200-300GB per dataset
- Tumor transcriptome profiling from the OAK (Rittmeyer *et al*. The Lancet. 2017) and POPLAR (Fehrenbacher et al. The Lancet, 2016)randomized clinical trials of atezolizumab in 2L NSCLC
  - Raw data available at EGA: **EGAC00001002120**
  - https://ega-archive.org/datasets/EGAD00001007703
  - Data type : .fastq file
  - Size: +- 500GB

Guidance:

Data can be digital or physical (for example biobank, biological samples, …). Data type: Data are often grouped by type (observational, experimental etc.), format and/or collection/generation method.

Examples of data types: observational (e.g. survey results, sensor readings, sensory observations); experimental (e.g. microscopy, spectroscopy, chromatograms, gene sequences); compiled/aggregated data[5] (e.g. text & data mining, derived variables, 3D modelling); simulation data (e.g. climate models); software, etc.

Examples of data formats: tabular data (.por,. spss, structured text or mark-up file XML, .tab, .csv), textual data (.rtf, .xml, .txt), geospatial data (.dwg,. GML, ..), image data, audio data, video data, documentation & computational script.

digital data volume: Please estimate the upper limit of the volume of the data per dataset or data type.

physical volume: Please estimate the physical volume of the research materials (for example the number of relevant biological samples that need to be stored and preserved during the project and/or after).

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.	
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, please describe these issues further and refer to specific datasets or data types when appropriate.	☒ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ☐ No If yes, please describe: - Reference to ethical committee approval: S63531, approved by the Ethical Committee of UZ/KU

[5] These data are generated by combining multiple existing datasets.

	Leuven - We will generate genomic, (single-cell) transcriptomic and (single-cell) proteomic sequencing data of both tumour samples and serial blood samples of patients treated for advanced lung carcinoma at the Department of Respiratory Oncology in UZ Leuven. All patient samples are labelled with study sample IDs, while the coding key remains with the treating oncologist at UZ Leuven. The coding key does not carry any personal identifiers and all records containing the identity of each participant is kept private and confidential. Access to the coding key is necessary to link data or biological samples back to a subject identifier and will be done under supervision of Prof. Dr. Els Wauters. - Clinical patient information will collected, linked to the study sampleIDs and pseudonymized before registration in a customized RedCap database, under supervision of Prof. Dr. Els Wauters. Basic demographic data are recorded, including age, gender, concomitant diseases, concomitant medication, time and stage of first diagnosis, location of the tumor, pathological and genetic diagnosis, performance status, previous therapy (type, duration and response, if applicable) and stage of disease at study entry. Clinical outcome parameters (according to the mRECIST criteria) are also recorded, including disease control rate (stable disease or response) and objective response rate (partial or complete response). Finally, survival data is collected (PFS and OS). - Sequencing data generated within this projected will be correlated with the pseudonymized clinical data (e.g. response data, survival data etc.). The sequencing data and associated pseudonymized patient information are defined as sensitive personal data and will 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. - Data collection details and the strategy to guarantee the privacy of the study participants are specified in the research protocol approved by the Ethical Committee of UZ/KU Leuven (reference

	number S63531). Prior to inclusion, each study participant is required to give consent for the processing of personal data for the purpose of the current study, using the most recently updated 'Informed Consent document' (ICF) and 'Patient Information Brochure' approved by the Ethical Committee of UZ/KU Leuven. All data is processed and stored on the institutional IT infrastructure, protected by a genuine user authentication system relying on username and password. Access to the data as well as the access level is limited on a project need and individual basis. Only researchers working on the project have access to these data. Due to the sample labelling as protective measure, the researchers are not able to decipher the identity of the donor.
Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☒ Yes ☐ No If yes: - Short description of the kind of personal data that will be used: Cfr. supra - Privacy Registry Reference: The UZ Leuven KWS system will serve as the source for the clinical information and electronic CRFs will be used for collection of these coded data. All obtained research data points and clinical information will be added in a coded manner to a REDCap database, located on a KU Leuven hosted and secured server which is double password protected. A dedicated, trained person will add all research information from this project to the database, which was especially designed for this research. Only coded information will be extracted and used for the downstream research analyses.

[6] See Glossary Flemish Standard Data Management Plan

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 ☐ No If yes, please comment: 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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☒ Yes ☐ No If yes, please explain: 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 Universitaire Ziekenhuizen Leuven / KU Leuven and VIB.

3. 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).	Human data at LTG will be processed in accordance with the institutional SOPs, the principles of the GDPR 2016/679 and the Belgian privacy law. All (human) samples arriving in LTG lab or processed/stored in LTG lab are labelled with a **study sample ID** (the coding key remains with the principal investigator/clinicians), keeping the identity of the study participant private and confidential. We only receive **pseudonymized** human data (including health data linked to the sample IDs) in .xls format. The sequencing data that we generate will be correlated with the pseudonymized personal data. After the arrival of the (human) samples at our laboratory, we will first save all sample information in our electronic **sample meta data sheet** (MTD) .xls format, containing all samples processed/stored at LTG. In this MTD, all the samples processed/stored at LTG will receive a **DILA ID** (double-coded) and this DILA ID is further used in the downstream analyses at LTG. When the data analyses are complete and ready to send, the DILA ID is matched to the original sample ID before transferring the results.

Will a metadata standard be used to make it easier to **find and reuse the data**? If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse. *REPOSITORIES COULD ASK TO DELIVER METADATA IN A CERTAIN FORMAT, WITH SPECIFIED ONTOLOGIES AND VOCABULARIES, I.E. STANDARD LISTS WITH UNIQUE IDENTIFIERS.*	☒ Yes ☐ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: Sequencing data types require specific metadata when submitted to public repositories such as EGA, ArrayExpress, GEO or ENA. 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 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. If no, please specify (where appropriate per dataset or data type) which metadata will be created:

4. 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). All patient samples, ˜single-cell suspensions and sequence library preparations 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.
How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS? DESCRIBE THE LOCATIONS, STORAGE MEDIA AND PROCEDURES THAT WILL BE USED FOR STORING AND BACKING UP DIGITAL AND NON-DIGITAL DATA DURING RESEARCH.*[7] *REFER TO INSTITUTION-SPECIFIC POLICIES REGARDING BACKUP PROCEDURES WHEN APPROPRIATE.*	KU Leuven drives are backed-up automatically on a daily basis using KU Leuven services. All sequencing data stored on the Flemish Supercomputer Centre (VSC) will be regularly transferred to the archive area that is mirrored.

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

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 □ No If yes, please specify concisely: 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 eseries 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) If no, please specify: /
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *CLEARLY DESCRIBE THE MEASURES (IN TERMS OF PHYSICAL SECURITY, NETWORK SECURITY, AND SECURITY OF COMPUTER SYSTEMS AND FILES) THAT WILL BE TAKEN TO ENSURE THAT STORED AND TRANSFERRED DATA ARE SAFE.* [7]	Sequencing data and associated pseudonymized patient/clinical information is considered sensitive information and will be handled as such. All electronical data, including the health data, the sample info sheet, the MTD, the created sequencing data and data analyses, 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) which are protected by a genuine user authentication system relying on username and password. LTG has a state-of-the-art **computing infrastructure** in-house, and for high-performance computing LTG has access to the **Flemish Supercomputer Centre** (Vlaams Supercomputer Centrum, VSC). The responsible person in LTG for sequencing data handling is Bram Boeckx (bram.boeckx@kuleuven.be). He is an expert in computer science and has ample experience with sequencing data handling and working in the VSC environment. Access to the data as well as the access level will be limited on a project need and individual basis.

	Prior to transfer the results, (sequencing) data labelled with DILA IDs will be linked to the original study sample ID and transferred **via Belnet Filesender or secure copy**. **Double-coded raw and processed sequencing 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. Accession to this sequencing data will be made available to any individuals making a specific request and this request will be handled by the institutional data access committee. In the end the samples will be returned, stored at the UZ/KU Leuven biobank.
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 4 years of this FWO project is ~12,500 EUR. This estimation is based on the following costs: - The costs of digital data storage are as follows: €868,9/5 TB/Year for the “L-drive” and €519/TB/Year for the “J-drive”. - The cost of VSC archive is €70/TB/Year, and staging €130/TB/Year. - We expect costs to drop slightly during the coming four years. All costs for data storage will be covered by own funding

5. 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...).	All sequencing data collected in the scope of this FWO project will be retained for the expected 5 year period after the end of the project. All (remaining) biological samples are preserved for 50 years, in accordance with the guidelines of Biobank UZ/KU Leuven.
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 All biological samples are registered and stored at the Biobank UZ/KU Leuven, in accordance with their guidelines.
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 for 5 years after the end of the project is ~ €4350. This estimation is based on 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. All costs for data preservation will be covered by our own funding. Electricity costs for the freezers present in the labs are included in general lab costs.

6. Data Sharing and Reuse	
Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available. *Note that 'available' does not necessarily mean that the data set becomes openly available, conditions for access and use may apply. Availability in this question thus entails both open & restricted access. For more information: https://wiki.surfnet.nl/display/standards/info-eu-repo/#infoeurepo-AccessRights*	□ Yes, in an Open Access repository □ Yes, in a restricted access repository (after approval, institutional access only, …) □ No (closed access) ☒ 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 doublecoded 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 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. For VIB-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 appropriate Data Access Committee (DAC) and processed in consultation with the researchers that produced the 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 per dataset or data type where appropriate.	☒ Yes, privacy aspects ☒ Yes, intellectual property rights ☒ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☐ No If yes, please specify: 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. WE have a Material and Data Transfer Agreement (MTA) between the legal entities of KU Leuven (> UZ Leuven) and VIB; 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 Universitaire Ziekenhuizen Leuven / KU Leuven and VIB. 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-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-coded read count data matrix (linked to double-coded patient data) will be available on an interactive webserver (http://blueprint.lambrechtslab.org). - Double-coded patient data: Upon publication, all double-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 (co-)promoters' websites (https://gbiomed.kuleuven.be/english/research/50488876, https://www.vibcancer.be/diether-lambrechts) 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). - 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? *This could be a specific date (dd/mm/yyyy) or an indication such as ‘upon publication of research results’.*	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 do not need to be delayed – or ongoing projects requiring confidential data. In those cases, datasets will be made publicly available as soon as the embargo date is reached.
Which data usage licenses are you going to provide? If none, please explain why. *A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.* *Example Answer: E.g. “Data from the project that can be shared will be made available under a Creative Commons Attribution License (CC-BY 4.0), so that users have to give credit to the original data creators.”* [8]	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 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. For VIB-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 appropriate Data Access Committee (DAC) and processed in consultation with the researchers that produced the data. The DAC will provide general guidance in terms of policies and will be referred to in handling controversial cases.

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *INDICATE WHETHER YOU INTEND TO ADD A PERSISTENT AND UNIQUE IDENTIFIER IN ORDER TO IDENTIFY AND RETRIEVE THE DATA.*	☒ Yes □ No If 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. All data management costs will be covered by own funding.

7. 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 E-notebook that refer to specific datasets. The research and technical staff includes the PhD student(s), technical assistants and bio-informaticians directly involved with this research project.
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. More specifically, Gino Philips, a junior computer scientist will handle storage and back-up of the sequencing data, under supervision of senior computer scientist Bram Boeck who has extensive experience in data handling and use of the Flemish Super Computer (VSC) environment. Final responsibility for data storage & back-up lies with promotor of this project, supported by ICTS, HPC, gbiomed-IT staff and UZ-IT staff.

Who will manage data preservation and sharing?	The research and technical staff will ensure data preservation and sharing, with support from ICTS, gbiomed-IT staff, and UZ-IT staff. More specifically, Gino Philips, a junior computer scientist will handle data preservation and sharing of the sequencing data, under supervision of senior computer scientist Bram Boeck who has extensive experience in data handling and use of the Flemish Super Computer (VSC) environment. Final responsibility for ensuring data preservation and sharing lies with the promotor of this project, supported by ICTS, HPC, gbiomed-IT staff, and UZ-IT staff.
Who will update and implement this DMP?	The promotor of this FWO project carries the end responsibility for updating and implement this DMP.