# C3 DATA MANAGEMENT PLAN (DMP) 2022: VALIDATING THE APPLICATION OF A FIRST-IN-CLASS HUMAN SERUM-FUNCTIONAL IMMUNODYNAMIC STATUS (SFIS) ASSAY AS BONA FIDE IMMUNO-ONCOLOGY BIOMARKER

## DMP TITLE

## 1. ADMIN DETAILS

**Project Name:** VALIDATING THE APPLICATION OF A FIRST-IN-CLASS HUMAN SERUM-FUNCTIONAL IMMUNODYNAMIC STATUS (SFIS) ASSAY AS BONA FIDE IMMUNO-ONCOLOGY BIOMARKER
**Grant Reference No.:** C3/21/037
**Principal Investigator / Researcher:** Abhishek GARG
**Description:** Biomarkers linked to cancer patient's immunological status are pivotal for reliable decision making regarding therapy options and patient management. However, biomarkers that can capture dynamic functional immune-signaling in non-invasive and personalised manner are currently unavailable. Thus, we established a novel, 'first-in-class', serum-functional immunodynamic status (sFIS) assay, utilising an 'in sitro' (in situ/in vitro) approach to interrogate the serum "immunome". In ovarian cancer patients, sFIS assay successfully estimated the risk of malignancy and distinguished differential survival, hence exemplifying its co-diagnostic/co-prognostic utility. But, beyond this proof-of-concept, further analytical/ clinical validation of sFIS assays' broad utility across multiple cancer-types and its immuno-oncological application is critically required. Hence, in this study we aim to test sFIS assay's performance across sera from >800 patient across 5 different cancer-types, with or without anticancer interventions (surgery, chemo/radio/immuno-therapy). Together, these inquiries will greatly enhance the socio-economic impact of sFIS assay thereby creating bright valorisation and leverage effect opportunities.
**Institution:** KU Leuven

## 2. DATA DESCRIPTION

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

- Generate new data
- Reuse existing data

**Describe the origin, type and format of the data (per dataset) and its (estimated) volume, ideally per objective or WP of the project. You might consider using the table in the guidance.**

The research and technical staff will generate, collect, process, analyse and store the data listed below, as detailed in the project description. The description below applies to all the different WPs of the project, based on the different types of experiments already described in the project proposal. Our estimated volume expected to be occupied per WP of the project is around 1.5-3 GB per WP.

The following datasets will be generated:

1. General: The data in this project will be acquired on in vitro (immune) cell line models, with respect to serum screening, and two major immune pathways (NFkB signalling or interferon/IFN signalling) ultimately culminating into bioinformatics and biostatistical analyses. The main (non-analogue) datatypes derived would be images, numerical data, single cell data or frequency data. The main analogue data would consist of those pertaining to cell lines, immune models, experimental assays and (frozen/stored) protein/nucleic acid samples or cell lysates. I would like to specify that all the patient data used in this project will be originally derived from the respective oncologists with whom we will work at home (UZ Leuven) or abroad and hence we are required to arrange an ethical approval and 3rd party material-transfer agreements – all of which are currently already finalized (i.e., ethical approvals) or to be finalized soon (3rd party material-transfer agreements).

2. Experimental, simulation or reused publicly existing data

Dataset 2.1. – Digital images

Microscopy pictures, histology pictures/scans, gel scans, graphs, illustrations, figures.

Dataset 2.2. – Cytometry data

Flow Cytometry and fluorescence-activated cell sorting (FACS) data

Dataset 2.3. – Omics/Bioinformatics and patient's data

Genomics, transcriptomics and single-cell RNASeq data derived from public database like TCGA, GEO datasets, BROAD Single Cell Platform or originally procured from the clinical collaborators.

Dataset 2.4. – In vitro assay (machine-derived) data

Excel sheets with data derived from machines/plate readers meant for polymerase-chain reaction (PCR) assay (thermocycler machine), readers for colorimetric, fluorimetric or bioluminescence data, readers for cytokines data (e.g. Luminex, MSD reader).

Dataset 2.5. – Statistical data

Derived from statistical softwares used for processing above data e.g. GrahphPad Prism, R Data Studio, Plotly, Tableau, Morpheus (Broad Institute) or Microsoft Excel.

3. Derived and compiled data

Dataset 3.1 – Research documentation

Research documentation generated by the research and technical staff or collected from online sources and from collaborators, including ethical approval documents, readme files, laboratory notes, protocols.

Dataset 3.2 – Manuscripts

The results will be published as BioRxiv preprints and articles in peer reviewed journals.

Dataset 3.3 – Algorithms, softwares and scripts

Algorithms, softwares and scripts to investigate the single-cell data, in vitro assay data, and to integrate different omics layers will be designed. This includes, combined set of scripts, algorithms, visualization tools & computer programs.

Data will be 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), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf), bitmap (.bmp), .gif;
- Digital images in vector formats: scalable vector graphics (.svg), encapsulated postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai);
- Flow cytometry data: Flow Cytometry Standard (.fcs);
- Omics data from publicly-accessible sources: binary base call format (.bcl), .fastq(.gz), SOFT or MINIMAL file formats (GEO datasets).
- Structural variations data: .vcf(.gz), .bcf
- Read/UMI count data: .tsv(.gz), Matrix Market format (.mtx), .loom, .rds(.gz)
- Other software formats: GraphPad Prism outputs (.pzfx), .jtv, .gtr, .cdt, .atr, .tar, .wmf, .ape;
- Plate reader/Imaging software/Gel scanning software outputs: .pda, .scn, .svg, .xpt;

IMPORTANT NOTE: The above data formats apply differentially to various different data-types and are together part of the general data volume per WP. As such it is tough to specify at the start of project how these different file formats will apply to different data-types and estimate their volume; although any such information will be eventually available for further assessment. Of note, we are aware that certain softwares/pipelines create specific formats that cannot be accessed by alternative softwares/pipelines, hence all the raw data will always be available in the “Dataset 3.1 – Research documentation” files, for future reference or re-use.

## 3. LEGAL & ETHICAL ISSUES

**Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to the file in KU Leuven's Record of Processing Activities. Be aware that registering the fact that you process personal data is a legal obligation.**

- No

Not applicable.

**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)**

- Yes

Ethical approval already available from the Ethical Committee for Human research (EC) via participating clinicians/oncologists: S50887 (Prof. Sabine Tejpar); S64358 & S60753 (Prof. Maarten Lambrechts); S64035 (Prof. An Coosemans); S59207 (Prof. Dirk Timmermans/Prof. An Coosemans); S63773 & S56919 (Prof. Hans Wildiers/Dr. Sigrid Hatse/Prof. Christine DeSmedt); S57760 (Prof. Oliver Bechter); S63833 (Prof. Benoit Beuselinck); S60379 (Dutrelasco trial - Prof. Paul Clement).

**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?**

- Yes

The assay to be used in this research (i.e., sFIS assay) is already IP-protected via LRD, KU Leuven, and hence the data generated herein falls under the same IP restrictions. See here for more details: https://lrd.kuleuven.be/doc/immuno-stratification-top

**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?**

- No

## 4. DOCUMENTATION & METADATA

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

- Documentation will consist of notes in the electronic laboratory records (Excel sheet-based metadata files) that refer to specific experiments. Those notes will describe the biological samples used/generated, experimental setup and protocols used, results generated, the links to the specific computer location as well as the names of the respective datasets. We will also maintain a metadata sheet with the connection between lab samples and files on our data storage, so that data files, lab samples, and experimental notes remain properly linked. Detailed protocols will be written, including research methods, practices, instructions given to participants, etc., for each experimental initiative. This will be stored in Word or Excel files. Furthermore, a logbook will be kept in Excel containing all steps that were taken to develop the final methodology, date of implementation and name of the researcher who carried out the experiment. Algorithms, scripts and software usage will be documented and stored alongside the electronic laboratory records, e.g. GraphPad Prism. If new scripts, algorithms or software tools are finalized, they will be additionally described in manuscripts and on GitHub (https://github.com/CellStressImmunityLAB). Herein, additional support is also available from KU Leuven will regards to extra functionalities for GitHub: https://set.kuleuven.be/set-it/allservices/service-catalog/kuleuven-gitlab. Finally, we will also keep all the information (in dedicated Excel sheets, PDFs or Word-files) about purchased antibodies, cell-lines, mouse models and other analogue data-sources. Other relevant information about these reagents and tools (e.g., proof of antibody specificity) will be derived from initial

standardization and optimization experiments and will be retained along with general research documentation/meta-data files.

**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.**

- Yes

We will use various metadata standards as applicable for different experiment/datatypes, as already established elsewhere: https://fairsharing.org/. For instance, flow cytometry (https://flowrepository.org/), microscopy imaging (https://www.openmicroscopy.org/), qRT-PCR (http://miqe.gene-quantification.info/), and publicly available TCGA patient data analyses (https://gdc.cancer.gov/about-data/data-standards) have very well-defined pre-established meta-data standards. In case we do not have a metadata standard available for a technique/datatype, a metadata of the numerical datasets will be created manually (e.g. based on the Dublin core metadata standard). For most of the data, metadata will be provided as readme, word or excel files, containing all settings and technical descriptions of the experiment and data. For imaging data, a large part of the metadata is included in the header files of the original images. These files contain information regarding the acquisition settings. In parallel, detailed meta-data info will be integrated within the electronic laboratory records linked to each experiment (as described above).

## 5. DATA STORAGE & BACK UP DURING THE PROJECT

**Where will the data be stored?**

Digital files will be stored on KU Leuven data storage servers: All data generated during the project will be stored on the local KU Leuven servers, PI computers and backup hard drives. This will be initially located in the real-time folders (on lab provided laptops/PCs of the students or employees and local KU Leuven servers) and later only in the archive folders (archive is mirrored; on local KU Leuven servers, backup hard-drives as well as PI computers). Any algorithms, scripts or softwares originally generated during the project will be stored in private online git repositories of the PIs. As soon as the manuscript is publicly available, the repository will be changed to a public repository. Specific biological samples (e.g. cell lysates, serum/plasma samples) will be stored in a freezer (-20°C or -80°C) while cell lines will be stored in liquid nitrogen.

**How is back up of the data provided?**

The paper copies will be digitized and together with the digital data stored on the university's secure network drives with automatic daily back-up procedures (e.g. J-Drive for confidential data and KUL Enterprise Box for non-confidential data).

KU Leuven drives are backed-up according to the following scheme:

- 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 "KUL Enterprise Box" is backed up on all official lab PCs and laptops.

- Besides the above back-ups, further backups will be done on lab hard-drives thereby creating high data back-up security.

**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

Yes, currently we have > 1 TB free data-space at our disposal (KU Leuven network/online serves + lab-based backup hard-drives), which will be enough for the data as described in part 2 of this DMP. Nevertheless, if required we have enough expandability options available either for existing KU Leuven servers (i.e. J-drive) or new servers that can be rapidly procured (i.e. K- or L-Drive).

**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 500 EUR. This estimation is based on the following costs:

- The costs of digital data storage are as follows: approximately 52 EUR/100 GB/Year for the “J-drive” and approximately 10 EUR/100 GB/Year for the “KUL Enterprise Box”.
- Additional budget for data storage on backup hard-drives is also foreseen: ~200 EUR per 1TB backup hard-drive.

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

The “J-drive” and “KUL Enterprise Box” servers (as well as lab back-up hard-drives, password-protected) are accessible only by laboratory members and PIs (since they are connected to KU Leuven password protections and subject to direct access permissions from the PI) and are also mirrored in the archive folders (e.g. J: or I: drives).

## 6. DATA PRESERVATION AFTER THE 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 10 years after the end of the project will be applied to all datasets that will be maintained and backed-up both on KU Leuven network/storage drives as well as lab backup hard-drives. Cell lines will also be stored for at least or more than 10 years. Cell lysates or serum/plasma samples will be stored for at least 5 years unless the ethical procedure requires their immediate destruction after screening (beyond 5 years, the degradation of these samples makes their storage “useless” and would require repeating the experiments to re-generate the data).

**Where will the data be archived (= stored for the longer term)?**
Data will be archived on the secured university's network/storage drive, as described in part 5 of this DMP. Additionally, data will be stored offline on external lab backup hard drives when the project is finished.

**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 are 260 EUR. This estimation is based on 52 EUR/100 GB/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 upfront paid for from this C3 project.

## 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)?**

- No

**Which data will be made available after the end of the project?**
The full datasets will be made available after publication of the data and upon reasonable request with the PI. The PI in the present project is committed to publish research results to communicate them to peers and to a wide audience. Depending on the nature of some datasets, 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, that is hard to specify currently and can only be defined at the end of the project). Other data could either be published along with the manuscript or made available upon direct requests.

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

- In an Open Access repository (e.g. https://www.nature.com/sdata/)
- Upon request by mail
- Other (specify):

As a rule, datasets will be made openly accessible along with the manuscript in line with the scientific journals' data availability-policies. Sharing policies for some specific research outputs are detailed below:
- Digital datasets that support publications (including image, cytometry data, and simulation data) will be made publicly available as per the journals' data availability policy.
- 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 (daily logs, raw data) deposited in the electronic lab notebooks/metadata are accessible to the PI and the research staff, and will be made available upon reasonable 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, post-print pdf versions of publications will be made available in online repositories (e.g. ResearchGate and KU Leuven repository), 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. Publications will also be automatically added to our KUL institutional repository, Lirias 2.0.
- Algorithms, scripts and software: All the relevant algorithms, scripts and software code driving the project will be made available to public repositories such as www.github.com.
- Data that do not support publication will be either deposited in an open access repository or made available upon reasonable request by email.

**When will the data be made available?**

- Upon publication of the research results

As a rule, all research outputs will be made accessible, as per the policy of the scientific journal, at the time of publication. 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 - or ongoing projects requiring confidential data. In those cases, datasets will be made available as soon as the embargo date is reached.

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

Data will be always accessible within the lab as well as made available to researchers inside KU Leuven (open data). For data made available publicly (in line with scientific journals' data availability policy), there is no restriction on data access. For other data (e.g. those covered by creative commons-type licenses), researchers from outside KU Leuven may freely contact us with reasonable requests for data access/sharing and this will be made available to them for access in a secure fashion e.g. KUL Enterprise BOX data sharing links.

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

No costs are expected. If any occur, they will be covered by the requesting parties.

## 8. RESPONSIBILITIES

**Who will be responsible for data documentation & metadata?**

The PhD, technician and/or postdoctoral researcher associated with this project will be responsible for data documentation & metadata, under supervision of the PI.

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

Data management, storage and back up will be performed by the PhD, technician and postdoctoral researcher associated with this project, under supervision of the PI (Abhishek D Garg).

**Who will be responsible for ensuring data preservation and reuse ?**

The PI (Abhishek D Garg) will be responsible for ensuring data preservation and reuse.

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

The PI (Abhishek D Garg) bears the end responsibility of updating & implementing this DMP