IMI2 Project ID 101005077

CARE - Corona Accelerated R&D in Europe

WP8 – Management, ethics, communication, dissemination, and exploitation

D8.14 Final Data Management Plan

Lead contributor	15 – Scifeon ApS (SCIFEON)
Other contributors	02 - Janssen 01 - INSERM All WP Leads and LIPAC members

Document History

Version	Date	Description
V1.0	6 April 2021	First version of Data Management Plan
V2.0	10 June 2024	Second version of Data Management Plan
V3.0	31 March 2025	Final version of Data Management Plan

The CARE project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 101005077. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and BILL & MELINDA GATES FOUNDATION, GLOBAL HEALTH DRUG DISCOVERY INSTITUTE, UNIVERSITY OF DUNDEE.

Table of Contents

1 Introduction

This is the final version of the Data Management Plan (DMP) for CARE (Corona Accelerated R&D in Europe). Project progress has not necessitated any revisions to data management plan, except for the addition of a section on long-term storage and archiving (section 6.8).
The DMP follows the principles that research data are Findable, Accessible, Interoperable and Reusable (FAIR) as outlined by the European Commission Horizon2020 programme.

1.1 Definitions

- Beneficiary: a participant to the Project.
- CARE-4-DATA: data management system that is setup as part of Task 8.2.
- Consortium Agreement: the consortium agreement entered into for the Project by the Beneficiaries.
- Clinical Data: data related to or derived from a clinical study performed under the Project.
- Data: research data generated under the Project within the Project's objectives. For the avoidance of doubt, research data includes any scientific data, including preclinical, clinical, and regulatory data generated under the Project within the Project's objectives.
- Data Owner: the Beneficiary who owns specific Data in accordance with the terms of the Consortium Agreement.
- Data Owner Representative: user with the Data Owner role, as appointed by a Data Owner and being able to act on behalf of such Data Owner.
- Dataset: collection of certain Data.
- Digital Research Data: data qualifying as "digital research data" within the scope of Article 29.3 of Grant Agreement; i.e. after being digitalised, e.g. when an instrument output file is created or an Excel file with calculated results is saved.
- Dissemination: any public disclosure of Data by a Beneficiary using any appropriate means (other than public disclosure from seeking protection for and/or exploiting such Data), including but not limited to by means of scientific publication (in any medium), press release, on a website, or by presentation at a scientific conference.
- Grant Agreement: the grant agreement for the Project entered into between the Beneficiaries and the IMI2 JU.
- LabKey: the Integrated Research Data Management system deployed in INSERM-Bordeaux for managing clinical data and systems biology data.
- Project (or Action): the IMI2-CARE project.
- Public Repository: open access repository such as Zenodo or the European COVID-19 Data Portal.
- Upload: the process of sending data to CARE-4-DATA or LabKey system. Data Upload does not imply Dissemination.

2 Data Summary

This Data Management Plan (DMP) covers the overall flow of Data from the Action, from generation, decision on protection of Intellectual Property (IP) in accordance with the Grant Agreement and Consortium Agreement, up and until Dissemination of the Data through at least one online portal (if applicable), including the Data Repositories referred to in Section 3, which will be used to manage the Data and make them available online. This Section 2 provides more details on the different Data types expected to be generated during the Project. Section 5 provides more details on the Data flows within the Project and possible Disseminations of such Data.

The main objectives of CARE include:

1. To identify therapeutic candidates against the current SARS-CoV-2, and other potentially emerging SARS-CoV-2 clades, and related coronavirus genera,
2. To identify immune markers contributing to the host immune response to COVID-19 infection and their correlations with clinical and virological outcomes,
3. To pre-clinically assess ADME, PK/PD, potency, and safety of therapeutic candidates *in vitro* and in animal models, and
4. To advance lead candidates into Phase 1 and Phase 2 clinical trials in humans.

- **Objective 1: identify therapeutic candidates against the current SARS-CoV-2, and other potentially emerging SARS-CoV-2 clades, and related coronavirus genera (WP1- WP4)**

Three approaches will be implemented to identify therapeutic candidates:

- Repurposing of drugs designated for other diseases, by screening and profiling compound libraries contributed by partners, with the aim of rapidly progressing molecules to advanced stages of clinical testing,
- Discovery of small molecules, based on *in silico* screening and profiling of candidate compounds directed against SARS-CoV-2 and future coronavirus targets,
- Discovery, generation and characterization of virus-neutralizing antibodies targeting spike protein.

- **Objective 2: identify immune markers contributing to the host immune response to COVID-19 infection and their correlations with clinical and virological outcomes (WP5)**

The study aim is to understand the virus-cell interaction through OMICS-based approaches and to evaluate the viral and immune responses in COVID-19 patients. The integration of the results will allow the identification of biomarkers and immune correlates associated with treatment efficacy and disease prognosis. In addition, understanding the mechanism of action will open new avenues for antiviral therapy.

- **Objective 3: pre-clinically assess ADME, PK/PD, potency, and safety of therapeutic candidates *in vitro* and in animal models (WP6)**

Therapeutic candidates identified in Objective 1 will undergo a battery of tests to evaluate their pharmacokinetics, safety and efficacy *in vitro* and *in vivo*, in order to select one or more lead candidates to enter clinical trials. Moreover, *in silico* PBPK modelling

based on ADME data and other information will determine the doses to be tested first in animals then in humans.

- **Objective 4: advance lead candidates into Phase 1 and Phase 2 clinical trials in humans (WP7)**

Two phase 1 and one phase 2 clinical trials will be conducted in CARE. Phase 1 clinical trials will evaluate the safety, tolerability, PK and optimal dose of two lead candidates whereas phase 2 clinical trial will assess the clinical efficacy of one therapeutic candidate.

2.1 Types and formats of Data

Annex A provides a non-exhaustive overview of the types and amounts of Data expected to be generated in the Project to meet the Project objectives, as well as an indicative assessment of their IP protection.

2.1.1 Data format

Data will be generated in standard formats to improve broad accessibility. Raw Data formats will depend on the instrument used, which general commercially available. Processed Data and reports will be presented in commonly used formats such as XLS, XLSX, CSV, TXT, PPT, PDF, TIFF, fastQ, 3DX, CCP4, Molfile, FASTA, Genbank, PDB, sdf etc.

2.1.2 Use of existing data

During the project, any published or publicly available data that is relevant to the project will be used as much as possible. Publicly available data may originate from publications (peer-reviewed or preprint servers e.g. BioRxiv) and other public databases, and can include data on activity and structure of other antiviral small molecules or antibodies (which may be used as starting points for screening and profiling), protein structures and co-crystal structures for molecular modelling and rational drug design, viral mutations for drug resistance studies, data on virus-host interactions, pathophysiology, the innate and adaptive immune system and immune markers for input in the systems biology activities and clinical trial design, etc. However, it is expected that most of the data needed for the Action will be generated within the consortium itself.

2.1.3 Data utility

The Data generated during the CARE project may be useful for the CARE partners in the same or other WPs, for the broader scientific community as well as for pharmaceutical and biotech companies, national and international guideline consortia, public health authorities and European patient rights representatives. The Project will encourage sharing of Data as much as possible, in accordance with the principles of the Grant Agreement, the Consortium Agreement, and this Data Management Plan.

3 FAIR Data

3.1 Making data findable

Data from the Action includes various types of experimental and clinical Data related to specific compounds, and/or antibodies. The Action will establish a process to assign a unique CARE compound number (in addition to the original compound number from the providing Beneficiary) to assets progressed into the Action (precise stage-gate still to be assessed). All Data relating to compounds will then be linked to this CARE compound number.

Furthermore, all Data will be linked to specific SOPs or equivalent procedures/protocols describing how experiments are carried out, and how the Data are collected and processed, ensuring a high degree of reproducibility. These procedures/protocols will be submitted together with a link to the published Datasets.
Furthermore, relevant metadata will be recorded and published as part of the Datasets. The structure of these metadata will be submitted to the public repositories as parts of the Dataset.

As part of the Dissemination process, each new Dataset will be assigned a Digital Object Identifier (DOI), and all new or updated Datasets will be assigned a version number.
Thus, published Data should be findable through the following means:

- CARE compound number,
- Assay procedure number,
- Assay metadata, for example virus-strain target, or
- DOI and version.

3.2 Making data openly accessible

The selection procedure of which Data to Disseminate and the Dissemination timing is described in section 5.6.
If released and based on its type, Data will be Disseminated to certified repositories selected among:

- European COVID-19 Data Portal,
- Zenodo,
- Gates Open Research, or
- Other relevant EBI (European Bioinformatics Institute) public repositories.

Data will be Disseminated using standard file formats, allowing access using standard software tools.
Machine-readable license terms will be included in each Dataset.
For those subsets of Data which cannot be made publicly available upon generation a specific procedure according to article 29.3 of the Grant Agreement (third paragraph) is in place to ensure restricted special access, which is described under section 5.7.

3.3 Making data interoperable

Data will be published using standard file formats, allowing Data-exchange and re-use

between researchers at different institutions. Typical examples of such standard formats are:

- CSV
- PDF
- MOLFILE
- FASTA
- Genbank
- PDB
- MzXML for mass spectrometry data.

The choice of vocabularies will be handled specifically for different Data types and will be included in the description published with each Dataset.

3.4 Increase data re-use

Within IMI2-CARE, re-use of Digital Research Data is assured through the open access and special access rights mechanisms discussed in section 5 below, which include measures allowing third parties to access, mine, exploit, reproduce and (in the case of open access) disseminate the Digital Research Data free of charge.

In addition, for any type of Data, consortium partners and third parties can exercise access rights for research use under the Grant Agreement and the Consortium Agreement. A request thereto needs to be made within 5 years after completion of the Project, except for those access rights which are granted on royalty-free conditions per the Consortium Agreement (as they are granted perpetually).

4 Allocation of resources

4.1 Responsibilities for Data management

The responsibility for Data management is split into four parts:

- The staff of individual Beneficiaries are responsible for compiling and uploading data and are responsible for ensuring the quality and comprehensiveness of the data.
- Data owner representatives are responsible for managing access control and Dissemination.
- Scifeon is responsible for operating and administrating the CARE-4-DATA system.
- INSERM Bordeaux is responsible for operating and administrating the LabKey system.

4.2 Cost and potential value of long term preservation

Data will be uploaded to public data repositories in accordance with Section 5 below thereby ensuring long term preservation of said Data. The preservation associated cost may be covered by the repositories.

The potential value of long-term preservation lies in the broad access of the scientific community to the results thereby contributing to the better understanding and the development of vaccines and/or treatments for Corona viruses.

5 Repositories and Data Sharing

5.1 Data Repositories

System	Description, purpose and access
CARE-4-DATA	CARE-4-DATA serves as an internal database for Data. The system facilitates Data sharing within the consortium and administration of the Data Dissemination process. Data Uploaded to CARE-4-DATA are only accessible to the Data Owner until access to other Beneficiaries is explicitly granted by the Data Owner. The Data can only be Disseminated by an explicit decision by the Data Owner, in which case, the Data is Disseminated to a Public Repository identified by the Data Owner.
LabKey	The LabKey system deployed at INSERM-Bordeaux, is a biomedical research data management system for Clinical and systems biology Data, which will be made accessible to other Beneficiaries for hosting their Clinical, experimental and laboratory Data. The system facilitates Data sharing within the consortium. Data Uploaded to LabKey are only accessible to the Data Owner until access to other Beneficiaries is explicitly granted by the Owner.
Open access repositories	Dissemination is achieved by publishing the Data to a Public Repository, such as Zenodo or the European COVID-19 Data Portal.

5.2 CARE-4-DATA User Roles

Role	**Description**
Administrator	Administrators can see all Data in the system and coordinate the process of Data Owner's providing access rights. The persons in the following roles have administrator rights in the CARE-4-DATA system: Platform Lead at Scifeon, Senior Project Manager at Scifeon, and Founder & CEO at Scifeon.
Data Owner	Each Beneficiary can designate one or more users as Data Owner Representative who will be assigned the Data Owner's role in the system. These users have the following special rights, specifically regarding the Data owned by the Beneficiary they are representing:

	• Make the decision to Disseminate Data to a Public Repository or to apply for Special Access Rights. • Control sharing of Data with other Beneficiaries. • Request access for new users associated with their own Beneficiary.
User	Users can upload Data to the system, including compound registration and experimental results, and be designated as Data Owner in accordance with the above. The uploaded Data will be linked to the Beneficiary to whom the User is associated. A User is associated with a specific Beneficiary and can access Data owned by that Beneficiary. Based on participation in specific activities, a User can also access Data owned by other Beneficiaries if access is granted by the Data Owner.

5.3 Data sharing in general

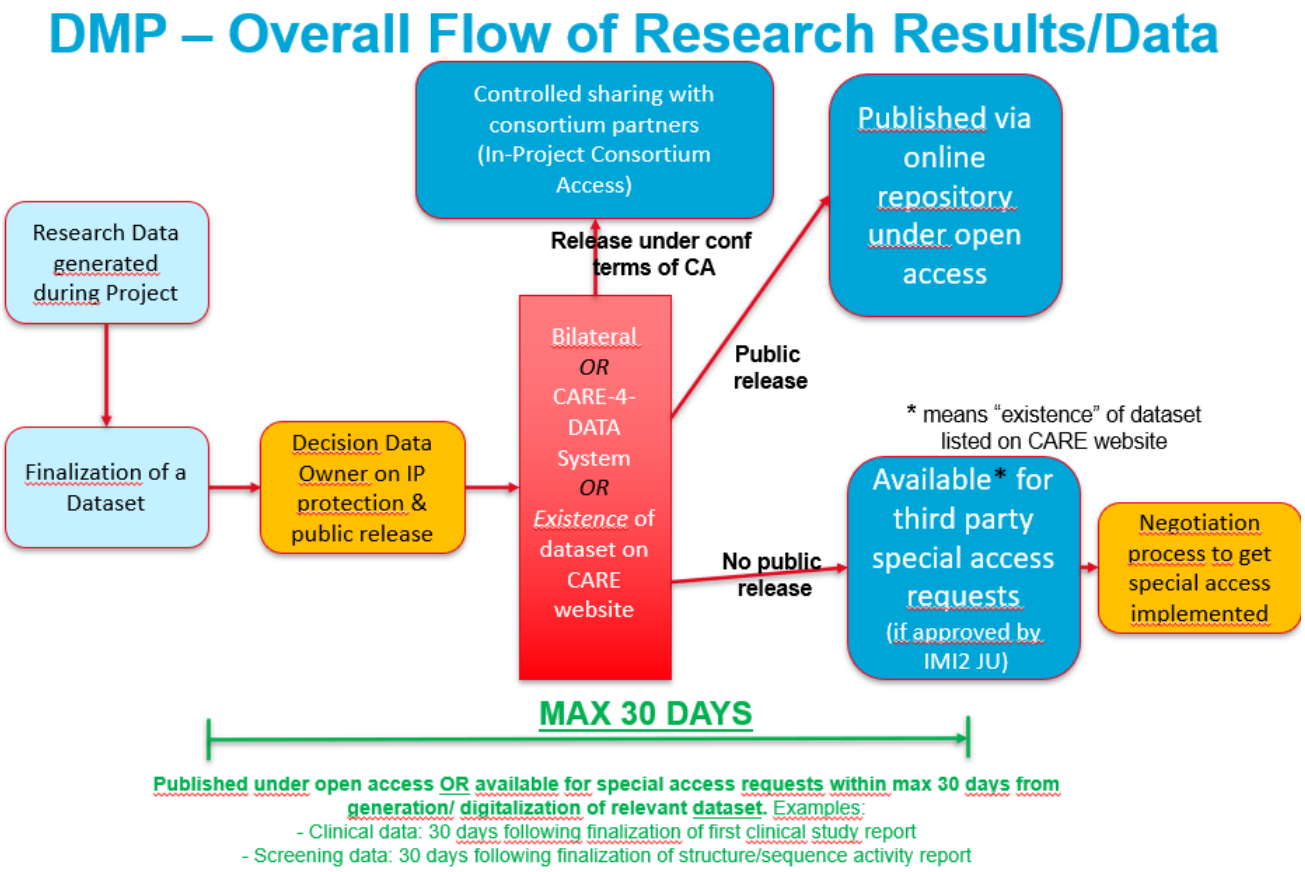
There are different mechanisms under which Data is made available:

1) Through access rights as provided for under the Consortium Agreement ("controlled sharing"),
2) Through Dissemination, including Open Access,
3) Through Special Access Rights.

In each case, the Data Owner shall retain full control on such Data sharing, subject to the terms of the Grant Agreement and the Consortium Agreement.

5.4 Data flow

This figure illustrates the flow of Data into and out of CARE-4-DATA:



Following generation of a Digital Research Dataset, the Data Owner shall assess within thirty (30) days following the generation/digitalization of the relevant Dataset whether the Digital Research Data in such Dataset requires IP protection and/or can be publicly released. Given that such assessment requires a full overview of available Data, the 30 days period does not start running from the generation/digitalization of each individual Digital Research Data point but only from generation/digitalization of the (full) relevant Dataset in which such Digital Research Data is to be included. For instance, (i) in case of clinical data the thirty (30) days' period starts running from the generation date of the first clinical study report; and (ii) in case of Data from compound/antibody screenings the thirty (30) days' period starts running from the generation date of the structure/sequence activity report. This contributes to ensuring that any further use of the Digital Research Data generates scientifically sound results and is not based on partially available or interim Data.

If a Data Owner decides that its Digital Research Data does not require IP protection, the Data Owner will upload the Digital Research Data in the CARE-4-DATA system and expressly indicate whether the Digital Research Data can be Disseminated and made available (i) under open access (see Section 5.6 below) or (ii) whether the special access rights regime should apply (see Section 5.7 below), although this may be less likely in

case no IP protection is needed. If the Data Owner so desires, the CARE-4-DATA system allows providing controlled access to the uploaded Digital Research Data to one or more Beneficiaries exercising their access rights in accordance with the Consortium Agreement.

If a Data Owner decides that its Digital Research Data does require IP protection, the Data Owner will undertake best efforts to obtain such protection within a reasonable timeframe. For patent applications, the Data Owner is expected to proceed with filing a first patent application within six months following the end of the aforementioned 30 days period. Reference is made to Annex A indicating eventual IP protection per category of Data expected to be generated in the Project. If the Data Owner so desires, the CARE-4-DATA system allows providing controlled access to such Digital Research Data to one or more Beneficiaries exercising their access rights in accordance with the Consortium Agreement.

Under the interpretation of Article 29.3 of the Grant Agreement of the IMI2 JU (see Annex B), while applying for IP protection, the Digital Research Data for which IP protection is sought shall be made available for special access rights requests from third parties within 30 days from generation/digitalization of the relevant Dataset. See Section 5.7 for details.

At the latest 30 days after sufficient steps have been taken to obtain IP protection for Digital Research Data, the Data Owner will upload the Digital Research Data in the CARE-4-DATA system and expressly indicate whether the Digital Research Data can be Disseminated and made available under open access (see Clause 5.6 below) or whether the special access rights regime should continue to apply (see Clause 5.7 below). For example, Digital Research Data protected via a filed patent application may be Disseminated and made available under open access at the latest 30 days after publication of the patent application if the special access rights regime does not continue to apply.

For the avoidance of doubt, the CARE-4-DATA system and any other system referred to in this DMP will not include any automatic process to release such Digital Research Data prior obtaining the Data Owner's express written approval or completion of an action by the Data Owner through the CARE-4-DATA system.

5.5 Granting Access to Data

Data can only be shared with other Beneficiaries through the CARE-4-DATA system by an explicit grant of read access by the Data Owner to that specific Beneficiary, said access can be revoked by the Data Owner at any time. Access must be granted within the CARE-4-DATA system by a user having the Data Owner Representative role. The grant of access occurs on a per-activity basis. When Data are Uploaded and tagged with a specific activity, Beneficiaries who have been granted access to that specific activity by the Data Owner Representative will immediately be able to read the Data and use it for the specific activity.

Other types of sharing, such as with third parties and through open access, always require the explicit approval of the Data Owner.

5.6 Dissemination

5.6.1 Dissemination Repositories

The Dissemination management module in CARE-4-DATA will support uploads to the repositories listed below. Data Owners can choose to use the system or select other ways of disseminating Data.

- Zenodo
- European COVID-19 Data Portal
- Gates Open Research
- Other relevant EBI repositories

5.6.2 Sustainability

Data sustainability beyond the duration of the Action is secured by uploading the Data to the repositories mentioned in Section 5.6.1.

5.6.3 Dissemination Process

With respect to the obligation to release the Digital Research Data under open access per Article 29.3 of the Grant Agreement, the Beneficiaries shall comply with Clause 7.5.4 of the Consortium Agreement.
For Datasets which are to be Disseminated and made available under open access in accordance with Section 5.4 above, the relevant Beneficiary first needs to provide a written notification to all other Beneficiaries in accordance with Article 29.1 of the Grant Agreement and Clause 7.5.2 of the Consortium Agreement. Following discussion with the IMI2 JU, a Consortium Agreement amendment will be initiated pursuant to which the Beneficiaries will be requested to amend Clause 7.5.2 of the Consortium Agreement to reduce the notice period from 45 days to 30 days.

5.7 Special Access Rights

Special access rights can only be applied for the Data types set forth in Annex A. Other Data types may also fall under the special access rights procedure, in which case Annex A will be updated.

If in accordance with Section 5.1 a Data Owner has flagged certain of its Datasets including Digital Research Data to be made available only under special access under confidentiality for the sole purpose of addressing the COVID-19 public health emergency, the identity number and a short description of the relevant Dataset as well as the identity of the Data Owner will be made public by the Data Owner (i) via an open access repository (see section 5.1) and through the CARE-4-DATA system; or (ii) via the CARE website. No other parts of the Digital Research Data will be made publicly available. The description will foresee that a third party interested in obtaining special access rights can request the Data Owner for additional information (only for the purposes of evaluating its interests in entering into a special access agreement) on the Dataset to evaluate its

interests accessing the Data. The additional information can be provided by the Data Owner following execution of a confidentiality agreement (CDA) with the requesting third party.

Thereafter, third parties may apply for special access rights to such Digital Research Data by written notice to the Consortium Executive Committee. The written notice must specify the specific Dataset that the request pertains to as well as the reasons for which such access is needed to address the COVID-19 pandemic.

After receiving the request, the Executive Committee will provide a reasoned non-binding advice to the Data Owner with respect to the request. The Data Owner will be invited to the relevant Executive Committee meeting during which the non-binding advice will be provided. Thereafter, the request needs to be considered by the Data Owner for confirmation whether the request is valid and whether access is needed to address the COVID-19 pandemic. The request will not be unreasonably refused by the Data Owner. Nonetheless, the Data Owner may refuse any third party's special access request if it pertains to commercially sensitive Data such as those identified in Annex A. Any negative decision will be forwarded to the IMI2 JU for information in a timely manner. Any positive decision needs to be reported to the IMI2 JU in the annual reporting to the IMI2 JU.
Before special access is granted, a special access agreement must be signed by the relevant Data Owner and the relevant Third Party. Any special access rights granted must include the right to access, mine, exploit and reproduce the data free of charge, but under confidentiality and must be limited (i) to the purpose of addressing the COVID-19 public health emergency and (ii) by any time limitation applicable to the special access right.

For clarity, a Data Owner can elect to provide special access rights only for the time period during which appropriate steps are undertaken to obtain IP protection on the relevant Dataset. Once IP protection is obtained, the Data Owner may elect to make available such Data under open access in accordance with Section 5.6. In such case, the relevant special access agreement with a third party being granted special access rights (see below) may include a time limitation.

5.8 Data in CARE-4-DATA

CARE-4-DATA is the shared data exchange platform for the Action, setup by Scifeon as part of Task 8.2. The following classes of Data will be stored in CARE-4-DATA:

- Compound and antibody registries
- Research results
- SOPs and protocols
- Study reports
- Public submission data packages from LabKey, containing clinical and omics data

5.8.1 Data Registration

For each Dataset, the following shall be provided via a standard template:

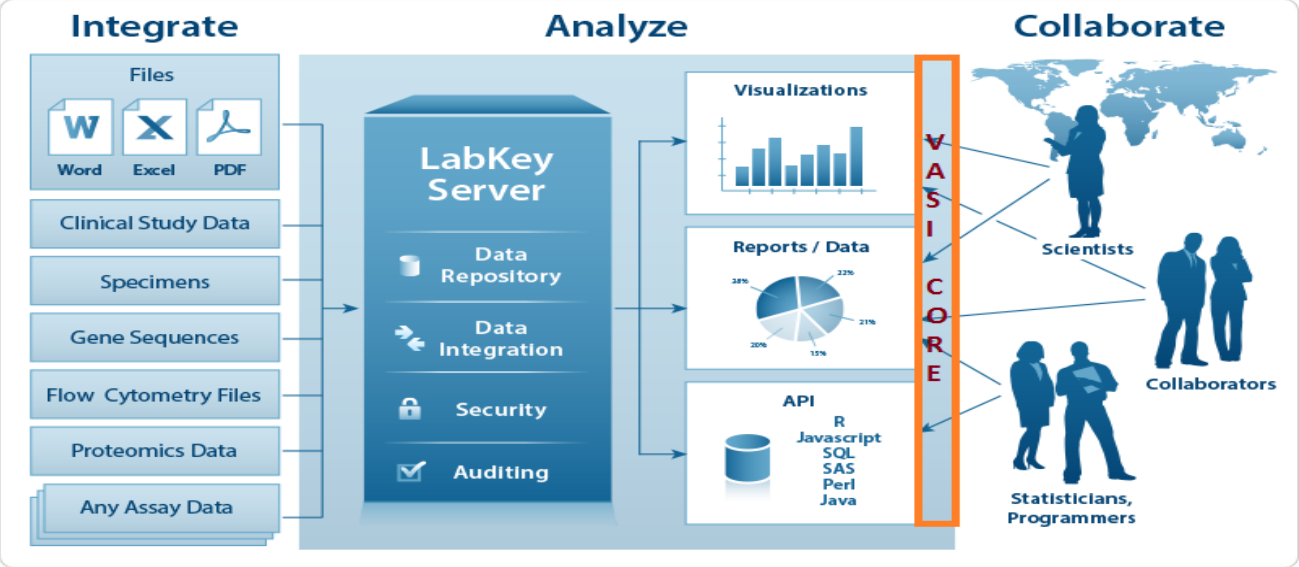
- Dataset reference and name
- Dataset description

- Standards and metadata
- Specific ethics and legal issues (*)
- Data protection, IP/copyright and ownership (*)

(*) To be aligned with Consortium Agreement and this DMP.
This template shall be completed when the Dataset is first imported and the whole set of completed Dataset templates shall be reviewed every six months.

5.9 Data in LabKey

5.9.1 Data in LabKey



LabKey Server brings Data together from different databases and multiple sources into one integrated repository for browse and analyze. Data can be of any shape and format, such as Data in Excel spreadsheets, instrument-derived assay data, patient questionnaires, medical histories, specimen inventories.

If required, anonymous and pseudonymous procedure can be applied to the uploaded data in LabKey server. Upload of pseudonymous data may require an additional data processing agreement if required by applicable laws.

- Before the reception of the Data, Platform Administrators will process and create the project folders and grant the right users access to the right Data following the Data Owner specific grant of access.
- Data can only be shared with other Beneficiaries through Labkey system by an explicit grant of access by the Data Owner to each specific Beneficiary, which can be revoked by the Data Owner at any time. Access to Data stored in Labkey repository follows the same rules defined in chapter 5.5 of this document.

Data submission:

- Data submitted to LabKey Platform must be done via a secured network:
  - The BPH Data Platform upload facility
  - Any other approved tool using encrypted connection for an administrator to upload on the Platform. This tool must encrypt data using at least 128 AES Encryption.

Clinical Data – Omics Data:

- Integrate clinical, demographic, and observational data from participants, cohorts and groups into LabKey Server
- Bring together Study Data along with related assay results and specimen Data from lab instruments
- Connect clinical, demographic, experimental Data and metadata to create a complete Dataset for analysis
- Organize all relevant Data in a central, secure environment for more efficient conservation, analysis, Data sharing and publishing through a web-based secure portal

5.10 LabKey Data Storage & Security

5.10.1 Data Security

Administrators of the Labkey Platform can read all Data in the system, create platform users and set project authorizations according to the access rights specified by the Data owner.

Labkey Platform Administrators in Inserm BPH are:Senior Project Manager and Data Architect at Inserm-Bordeaux, IT Director

Labkey provides Dataset-level and Folder-level permissions:
Folder-level security: The following permissions can be granted to a user for each project and/or sub-folder:

- Author: Data owners are Authors, they may add new data in the specified folders, update and delete only information they added.
- Readers: Readers cannot modify any data but have access to the specified folders with read only permission.
- Submitters: Submitters may submit new data, but may not read or change anything.

5.10.2 Study Dataset-level security:

The security for a study can be configured in addition to the typical permissions for a folder.
Study Dataset-level security gives a finer-grained control over access to the individual Datasets within a folder. Note that Dataset-level configurations override the Folder-level configurations. Thus it's possible for Readers of a folder to be granted Author permissions on selected Dataset in the same folder, or, on the contrary, to make selected Dataset unreadable to groups that otherwise have read access to the folder.

If required, Data exported from LabKey Server can be protected by:

- Special handling for PHI data (Protected Health Information data) such as mechanisms for anonymizing and obscuring participant ids, birth dates, exam dates, …

- Administrators can audit and review all of the activity pertaining to the secured data such as when users have logged in and where they have been inside the repository, what they have done.

- LabKey Server has a group- & role-based security model. Each user of the system belongs to one or more security groups, and each group has a specific set of permissions in relation to projects and folders on the system.

The following permissions can be granted to a user for each project and/or sub-folder by the Platform Administrators:
- Editor: Editors have read/write permissions on the folder but cannot manage permissions for users.
- Author: The author can create new Data, modify their own Data but cannot modify or delete Data from other Beneficiaries.
- Readers: Readers cannot modify any Data but (to the extent permitted by the respective Data Owners) have access to the Data in the platform with read only permission.

5.10.3 Data Storage

LabKey Server is a secure, web-accessible file repository, which serves both as a searchable storage place for files, and as a launching point for importing Data into the database.

While LabKey repository provides a storage, indexing and sharing location for unstructured data files, the server acts as a launching point for structured data files, to the database for more advanced analysis.
Search indexer scans and pulls out words and phrases to enable finding files with specific content.

For files that are opaque to the search indexer, such as biopsy image files, Labkey still serves as a staging point.

Data will be stored on servers hosted by the CREDIM, IT Platform Team of the INSERM accredited to host health related Data. These servers are situated in the ISPED building in Bordeaux, France.

A distant backup of the data is realized in another server room of the university, on specific server of the CREDIM and Inserm BPH.

Data will be backed by the CREDIM as indicated in the SER-SEP-POL-001 Information System Security Policy and the Business Continuity Planning.
Logins to the server and the Data held on the server using other means than the platform will only be granted by the CREDIM to a limited number of administrators.

5.11 FAIR Data in Labkey

Labkey addresses FAIR requirements for captured Data with persistent efforts to adhere to Findable, Accessible, Interoperable, Re-useable Data standards to the extent feasible. The appropriate research Data from WP5 and WP7 will be transferred to an Open access repository. Although the project aims to fulfil all the FAIR objectives set by GA and CA for making research results openly available within 30 days to the research community, we would also make sure that all the regulatory obligations (as set by the GDPR) as well as potential protection of intellectual property rights and any contractual obligation(s) can be fulfilled and the Data properly controlled and analyzed before being able to release them to open access.
The Data produced and/or used in the project including their metadata will be made identifiable and locatable by means of a standard identification mechanism.

5.11.1 Making data Findable

The Data produced and used in the project is discoverable, as both Data and metadata are fully indexed in Labkey platform. Data can be accessed and 'found' in many different ways, using queries, full text search, by creating visualizations and reports, by exporting in many different formats.
Data can be placed in a defined container or folder of choice in order for it to be easily and directly available.
Data can also be made publicly accessible, using the PUBLISH tool, so that internal and external research partners with the permission right and even the wider public can have access to the appropriate Data.

5.11.2 Making data Accessible

Data is accessible in Labkey by deposition in the Data repository where associated metadata will reside. Users can be given permission to defined users and user groups, specific projects, folders, Data sets and specific tables, depending on what is appropriate and defined by Data manager of the project and the platform's administrator.
Data can be presented in dashboards, any one of a number of visualizations in order to make it more accessible.

5.11.3 Making Data Interoperable

Data in Labkey repository requires no additional software else than a secure web access. LabKey is a Data integration tool, allowing Data exchange and re-use between researchers, facilitating re-combinations with different datasets from different origins. Furthermore Labkey has many tools that allows Data and tables to be joined in a way that makes previously unusable Data, usable. Using foreign keys and lookups for example, LabKey add context to Data and these processes are often automated, or in some way triggered or can be done on a manual basis. Data extracts from Labkey can be obtained in the following formats; Excel, CSV, and TXT formats that can easily be imported into many other tools.

5.11.4 Making data Re-usable

Only Data owners can guarantee the Data is reproducible. Labkey provides a set of principles (previously mentioned) and validation tools to allow users to confine Data, use limits, regular expressions to ensure the Data remains clean.

6 System Security

6.1 CARE-4-DATA System Security

- All Data uploaded to CARE-4-DATA are stored in Microsoft Azure databases.
- All Data are under audit trail, so that no Data can be deleted from the database and all Data modifications are tracked.

The security focus for CARE-4-DATA is on minimizing the risk of accidental publication of IP sensitive data. This is achieved by:

- Except if otherwise agreed by the Data Owner, compound structures or protein sequences which are pre-existing or which have been generated outside of the Project will not be uploaded into the system and will only be stored by the compound or the protein owner.
- Controlling data access as is set forth above.
- Operating the system based on a set of standard elements (web server, database server) hosted by Microsoft in an Azure environment, as described below.

Other key information security aspects are:

- Ensuring the integrity of the Data that is being shared between Beneficiaries.
- Ensuring high availability of the system.

This section describes the general information security management for CARE-4-DATA.

6.1.1 Access Control

Each Data Owner nominates a Data Owner Representative who can authorize creation of user accounts to access Data owned by said Data Owner. Furthermore, the Data Owner Representative can authorize controlled sharing of individual results with other Beneficiaries.

6.2 CARE-4-DATA System description

CARE-4-DATA consists of two Scifeon cloud environments: a test-environment located at https://care4data-test.scifeon.cloud and a production-environment located at https://care4data.scifeon.cloud.
Scifeon environments are *completely* separated (i.e. single-tenant) from each other and always consists of the following services:

- A HTTP server with all endpoints for the Scifeon browser client
- A SQL Server Database
- A file storage service
- A browser client served from the HTTP server and CDN's managed by Scifeon

The different environment types are never hosted on in the same databases or HTTP servers.

6.2.1 Traffic

Access to storage and database is limited to internal communication between services at the cloud provider, or if necessary, made public for only specific IP-addresses.

All communication to the Scifeon HTTP server must be encrypted using the HTTPS-protocol. HTTP is disabled, and all traffic using HTTP will be redirected to HTTPS. The cryptographic protocol TLS version 1.2 or above is required to communicate with the Scifeon HTTP server.
All internal communication to storage services and SQL databases must also be secure (ensured by Azure).

6.2.2 Data storage

The Scifeon HTTP server and database are both located in the Azure region West Europe within the geographical location Europe. This region is located in Netherlands.
Microsoft guarantees all data stays within a defined geographical region:
"Microsoft may copy customer data between regions within a given geo for data redundancy or other operational purposes." https://azure.microsoft.com/en-us/global-infrastructure/data-residency/#more-information
All data stored in the Scifeon database and file storage is encrypted at rest (https://docs.microsoft.com/en-us/azure/security/fundamentals/encryption-atrest). On request Scifeon can use a user's own encryption keys for encrypting data at rest, i.e. data in databases and files uploaded. Scifeon is using standard Azure functionality for enabling this feature: https://docs.microsoft.com/en-us/azure/azure-sql/database/transparent-data-encryption-byok-overview?view=azuresqldb-current

6.3 Labkey System and Data Security description

- Labkey platform deployed at Inserm-Bordeaux organizes all relevant data in a central, secure environment for more efficient conservation, analysis, data sharing and publishing through a web-based secure portal.

Labkey's secure environment is composed of:

- A Web server with:
  - Encrypted HTTPS-protocol for all communications via the server.
  - SSL/TLS Certificates on Tomcat server
  - The Tomcat server uses a Java KeyStore (JKS) repository to hold all of the security certificates and their corresponding private keys.
- A file repository server: LabKey data are stored on a dedicated directory on a Linux cluster of two physical file servers connected to a specific storage bay.
- A SQL Server Database: Labkey database is stored on a cluster of two physical Microsoft SQL/Server connected to a separated specific bay.

The LabKey web server is hosted in the DMZ area of the Firewall and protected with FortiNet VPN. The LabKey File Repository and the SQL/Server database are protected into the LAN area of the firewall.

All Data must be submitted to LabKey Platform via a secured network: the BPH Data Platform upload facility (FTPs server) or by uploading data directly to the secured Labkey Data repository.
Any other approved tool using encrypted connection for an administrator to upload on the Platform. This tool must encrypt data using at least 256 AES Encryption.

In addition, as a first step prior to the transfer of the data of the EPIC-HR cohort from Pfizer to Inserm-SISTM the LabKey environment security has been tested. This was important for Inserm-SISTM, which had recently carried out an upgrade of its datawarehouse environment to the NIS2 standard following European directives on cybersecurity. Following this implementation and in order to respect Pfizer security policies, a number of additional tests were run in order to check that everything would be up to the highest standards. Following the test, it was decided to strengthen security measures for sensitive data with an up to date solution and the adaptation of the CSP. This was solved by implementing a dedicated VPN and a SSO system. With this solution able to secure access even for further development, the security of the data on the LabKey solution is thus now up to the current highest security levels. Furthermore the the existing GLPI for automated and centralized management CMDB (Configuration Management Data Base) implementation has been completed with complete information of the IT inventory, including asset ownership and operational status.

6.3.1 Full Audit trail

- Events that happen anywhere on an instance of LabKey Server are logged to track down issues or document what has occurred for compliance purposes.

Administrators can audit and review all of the activity pertaining to the secured data such as when users have logged in and where they have been inside the repository, what they have done. Administrator can set the level of auditing detail on a table-by-table basis, determining the level of auditing for insert, update, and delete operations. Different types of events are stored in separate log categories on the master **Audit Log**. The categories for logged events are:

- **Assay/Experiment events**: Assay run import and deletion, assay publishing and recall.
- **Attachment events**: Adding, deleting, and downloading attachments on wiki pages and issues.
- **Authentication settings events**: Information about modifications to authentication configurations and global authentication settings.
- **Client API Actions**: Information about audit events created by client API calls.
- **Copy-to-Study Assay events**: Events related to copying assay data into a study.
- **Dataset events**: Inserting, updating, and deleting dataset records. QC state changes.
- **Domain events**: Data about domain (list, dataset, etc.) creation, deletion, and modification.
- **Domain property events**: Records per-property changes to any domain.
- **File batch events**: Processing batches of files.
- **File events**: Changes to a file repository.
- **Flow events**: Keyword changes in the flow module.
- **Group events :** Any action Administrator performs relative to the groups of users
- **Inventory Events**: Events related to freezer inventory locations, boxes, and items.
- **List events**: Creating and deleting lists. Inserting, updating, and deleting records in lists.
- **Logged query events** - Shows the SQL query that was submitted to the database
- Many other events …

6.4 Hosting infrastructure

Scifeon is a SaaS offering hosted in the Microsoft cloud-environment Azure. Each of the components outlined above are hosted using the following managed services:

- HTTP server: App Service (https://azure.microsoft.com/en-us/services/app-service/)
- SQL Server Database: SQL Database (https://azure.microsoft.com/en-us/services/sql-database/)
- File storage service: Blob storage in Storage Accounts (https://azure.microsoft.com/en-us/services/storage/blobs/)
- Browser client: stored in a Blob Storage served as a CDN (https://docs.microsoft.com/en-us/azure/cdn/cdn-create-a-storage-account-with-cdn)

As a company Scifeon doesn't own or run any virtual machines, nor internal or customer facing. Scifeon's Active Directory is hosted in Azure, Scifeon uses Sharepoint for documents and Office365 is used for emails. Task tracking is also done using cloud services.

6.4.1 Microsoft responsibilities

Microsoft does network penetration testing of their data centers: https://gallery.technet.microsoft.com/Cloud-Red-Teaming-b837392e
The physical security of their data centers are described:

> "Microsoft designs, builds, and operates datacenters in a way that strictly controls physical access to the areas where your data is stored. Microsoft has hundreds of Azure datacenters in 54 regions (as of 2019), and each of these has extensive multilayered protections to ensure unauthorized users cannot gain physical access to your customer data. Layered physical security measures at Microsoft datacenters include access approval:
>
> - At the facility's perimeter.
> - At the building's perimeter.
> - Inside the building.
> - On the datacenter floor.
>
> Physical security reviews of the facilities are conducted periodically to ensure the datacenters properly address Azure security requirements."

Trusted Cloud eBook: https://go.microsoft.com/fwlink/?LinkId=392408&clcid=0x409

6.4.2 Information and data

Cloud environments are created using automated pipelines and templates, which follow all Azure security recommendations.
The Azure services used have certain security baseline recommendations described here:

- Azure SQL Database: https://docs.microsoft.com/en-us/azure/azure-sql/database/security-baseline
- Azure Storage: https://docs.microsoft.com/en-us/azure/storage/common/security-baseline
- Azure App Service: https://docs.microsoft.com/en-us/azure/web-application-firewall/security-baseline

Application penetration testing and vulnerabity scans are periodically performed by third-party for Scifeon. These reports are made public if requested.

6.4.3 Devices (Mobile and PCs)

Scifeon offices are locked with chip+pin and keys to individual offices. Computers are always locked when left unsupervised, and passwords are at least 15 characters long.

6.4.4 Accounts and identities

All access to relevant information security management systems in Azure is controlled through AD roles in the internal Scifeon AzureAD.
Scifeon personel who can access audit logs in production environments are restricted using AzureAD groups.

6.5 Source code & deployments

Scifeon DevOps pipelines log all deployments to all environments, including production. The logging includes who initiated the deployment, when it was done, which components were updated during the deployment, the version of the deployment, and if everything were successfully deployed.
Installing and uninstalling apps are also logged with the same level of detail as general deployments.

6.5.1 Accounts and identities

All access to source code of Scifeon is located in Git repositories in Azure DevOps. Azure DevOps access is controlled by AzureAD groups.

6.6 Authentication and Access Tokens

Authentication is typically initiated through the Scifeon browser client and the calls to the HTTP server.
Successful sign-ins generate an accessToken (valid for 30 minutes) used for authenticating to the Scifeon HTTP server. A refreshToken (no expiration unless "Keep me signed in" is not checked, then it expires after 8 hours) is also generated, this is used for generating new accessTokens when these are expired. On each regeneration of the accessToken, the refreshToken is also regenerated, if it has an expiration.
Using the Scifeon client in the browser the accessToken is automatically regenerated every 20 minutes, i.e. before it is actually expired. This is to ensure all requests to the HTTP server is authenticated successfully.
The accessToken is provided in the header of each HTTP request and is validated using standard libraries, as described here: https://devblogs.microsoft.com/aspnet/jwt-validation-and-authorization-in-asp-net-core/
Tokens are signed using the standard algorithm HMAC+SHA256 and a unique key for each Scifeon system. Tokens are validated on each HTTP request.

6.6.1 Password based

Users can authenticate using a username (typically email or initials) and password. The password must be at least 8 characters long. When new users are added to the system, they are asked to create a new password immediately.
After 3 unsuccessful sign-ins to the same user account, i.e. using the same username, the account is locked. The account can only be unlocked by requesting a password reset, which will send an email with instructions on how to reset the password.

6.6.2 Azure AD based

Users can authenticate using their AzureAD (or better known as Microsoft or Office365) login. The first time a user uses this method, they must accept to allow Scifeon to access their AzureAD.

6.7 Backup

Data is stored in databases and file storage services; these are regularly backed up as described below.

6.7.1 Files

Files are geo-replicated in Europe, so if the primary data-center fails, the secondary data-center is used automatically:
https://docs.microsoft.com/en-us/azure/storage/common/storage-disaster-recovery-guidance

6.7.2 Databases

Point-in-time backups are available for the last week. Backups are stored in geo-redundant (RA-GRS) storage blobs that are replicated to a paired region: https://docs.microsoft.com/en-us/azure/azure-sql/database/automated-backups-overview?tabs=single-database
Long-term retention backups are enabled per customer.

6.8 Long term storage and availability

Scifeon will keep the CARE-4-DATA system running and fully available for a duration of 1 year after the project is completed on April 1st 2025.

Furthermore, Scifeon will keep the CARE-4-DATA system running in read-only mode for another three years.

7 Ethical aspects

The CARE consortium members adhere to all relevant international and national laws and guidelines relating to the conduct of the activities in the project. The research will comply with the highest ethical standards.

For all work involving human data collections, electronic or otherwise and sharing of personal data, CARE adhere to Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), as well as all other applicable national laws and regulations.

All research activities within CARE requiring approval on ethical and legal grounds through responsible local or national Ethics Committees and Regulatory Authorities, will be conducted only after obtaining such approval.

Patient Information and Informed consent procedures in relation to clinical studies to be performed under the project will be approved by the relevant national or local ethics boards or other competent authorities. Only patients able to give their consent will be included in any studies performed under the project.
There are no other ethic topics currently identified beyond those mentioned above and in Section 5 of the project's description of the Action (Annex 1 of the Grant Agreement).

Annex A

IMI-2 CARE – Expected types of data and indicative assessment of IP protection

Relation to the project objectives	Data-generating activity	Examples of types of data generated	Amount of data generated	Likely subject to IP protection Y/N (*)
Obj 1. Screening and profiling of chemical compounds and antibodies	Biological assays such as a phenotypic infectious virus-cell-based high-throughput screening assay, in vitro toxicity assays, biological validation assays and broad-spectrum assays.	*In vitro* activity and toxicity data (e.g., EC_{50}, IC_{50}, CC_{50} K_D values), RT-qPCR and virus titration data	It is foreseen that 400 – 500,000 compounds and about 30.000 siRNAs will be assayed	- Yes for compounds and antibodies showing anti-viral activity and/or target binding. This is considered commercially sensitive Data - No for compounds and antibodies showing no anti-viral activity and/or no target binding
	Target-based assays such as biochemical assays on the main protease, the RNA polymerase, the endonuclease, the exonuclease, or the two methyltransferases	Functional (enzyme activity and ligand-binding), biophysical (Tm, Biacore), biochemical, structural (X-Ray or Cryo-EM) data	It is foreseen that about 5000 compounds will be assayed	
Obj 1. Generation of further starting points for screening and profiling	In silico artificial-intelligence (AI)-based virtual screening	New small molecule chemistry	It is foreseen that about 500,000 compounds will be assayed	- Yes for compounds and antibodies showing anti-viral activity. This is considered commercially sensitive Data - No for compounds and antibodies showing no anti-viral activity
Obj 1. Generation and	Methods applied include the use of fully human phage and	Binding kinetics (such as k_a, k_d, K_D), in vitro potency	It is foreseen that about 1600	- Yes for antibodies showing binding

characterization of antiviral antibody leads	yeast display platforms, immunisation of humanised mouse or chicken models, in silico designed antibody libraries and isolation of protective antibodies from human patient serum and B cells	(IC_{50}), biophysical (Tm, Biacore), CMC (expression titers, Protein A recoveries, % monomer, thermal stability), biochemical, structural (X-Ray or Cryo-EM) data	antibodies will be assayed	properties. This is considered commercially sensitive Data - No for antibodies with no binding properties
Obj 1. Mode of action (MoA) elucidation	Target-based assays such as mentioned above, time-of-drug-addition (TOA) studies, selection of drug-resistant SARS-CoV-2, ex vivo models such as ALI culture systems and primary differentiated cell cultures	Antiviral activity data. For antibodies, in addition CDC, ADCC and ADCP, isotyping data	It is foreseen that about 100 molecules will be assayed	No
Obj 1. Optimization of hit properties to come to lead candidates for *in vivo* development	Assays such as the ones mentioned above to optimize potency, cell permeability, selectivity against off-targets and drug-like PK and physicochemical properties where possible complemented by molecular modelling and rational drug design	Structural and activity data as above	It is foreseen that about 100 to 500 compounds will be assayed. The SAR alone may yield a data volume of 10s to low 100s of Gigabytes	- Yes for compounds showing anti-viral activity and/or improved potency. This is considered commercially sensitive Data - No for compounds and antibodies showing no anti-viral activity and/or no improved potency
Obj 2. Understanding virus-host interactions	Study of virus-cell host interactions through an integrated OMICS approach (genomics, transcriptomics,	Transcriptional profile of infected HAE (Human reconstituted airway epithelia) and list of deregulated genes during	Transcriptomics: data from at least 10 HAE from different donors; Proteomics: data	No

	proteomics & metabolomics) in infected cells	the time course of infection, proteomics data of SARS-CoV-2 infected cells (HAE), phosphoproteomics and analysis of other post-translational modifications, mass spectrometry data of (phosho)proteome and metabolome in coronavirus-infected cells	from 4-8 time points post infection for a dozen post translational modifications, hundreds to thousands of protein-protein interaction descriptions; Genomics and Metabolomic: data from 4 different coronaviruses	
Obj 2. Characterization of innate and adaptive immune cells	Analysis of phenotypic profile and phosphosignaling of innate and adaptive immune cells (flow cytometry, Luminex assay, Illumina, etc.), characterization of T and B cell-specific responses (flow cytometry, ELISPOT assay, mass spectrometry)	Phenotype of immune cells, seric cytokine measurement, gene expression profiling (whole blood), specific T & B cell responses	It is foreseen that samples from 150 patients from COVID-19 patients and 180 patients from CARE clinical trials will be analyzed	No
Obj2. Identification of immune markers contributing to the host immune responses	Statistical data integration methods and mathematic modelling	Integrated analysis data	Size will vary in function of the number of patients analyzed	No

Obj 3. Pre-clinical evaluation of selected lead molecules in relevant animal models	Studies include ADME, PK/PD, efficacy, potency, non-GLP and GLP toxicity studies. Animal models include transgenic mouse models expressing human receptor ACE2, Syrian hamster models and non-human primate (NHP) models	*In vivo* toxicity, efficacy, safety and ADME/PK data such as $C_{through}$, C_{max}, AUC, maximum tolerated dose, potential side effects or mortality etc, virus shedding using qPCR	*In vivo* efficacy data in small rodents for up to 35 molecules, efficacy in non-human primates for up to 5 molecules	Yes. This is considered commercially sensitive Data
Obj 4. Clinical efficacy and safety evaluation of therapeutic candidates in Phase 1 and Phase 2 clinical studies	Clinical studies and patient sample analysis	Clinical patient data, data resulting from specific immunological and virological laboratory assessments	Two Phase 1 studies and one Phase 2 study which will generate about 5 – 10 GB of data	Yes. This may be considered commercially sensitive Data

(*) final decision on whether IP protection is needed or not, and whether said protection will trigger a delay in providing open access, always remains with the owner of the data

Annex B

IMI2 JU Interpretation of Article 29.3(1c) IMI2 JU MGA

In relation to the interpretation of Article 29.3(1c) IMI2 JU MGA and the obligations which are established therein, following consultation of the relevant Commission services, the IMI Programme Office confirmed the following:

the rationale of the specific provision (and by extension of the action) is to make all relevant data needed for tackling the public health emergency available (in principle in open access but providing for an exception where appropriate) as quickly as possible (at the latest within 30 days) as time is of the essence in case of addressing a public emergency and sharing of data is a key element). At the same time, the provision specifies that certain GA obligations should however not be disregarded, including the obligation to protect results in Article 27.

Furthermore, the provision sets out an exception to providing open access, i.e., through providing special access rights. Under such special access the same rights - i.e., to access, mine, exploit ad reproduce the data free of charge - must be provided except for the right to disseminate. Therefore, the intent of the exception is to provide for an alternative where dissemination (making data public) would be problematic, in particular in view of the GA obligations. In this respect, when prior dissemination in the public domain could indeed be problematic in view of obtaining protection, it can be replaced by special access rights. The latter would be the subject to a requirement to keep the data confidential until protection is achieved. The exception should be used until the moment that the dissemination would no longer constitute a problem and thereafter be replaced by open access.

In relation to the specific point raised by the CARE consortium, it is confirmed that that the exception does not need to repeat "at the latest within 30 days after it has been generated". As the scope of the exception concerns the rights which need to be given - i.e., to access, mine, exploit ad reproduce the data free of charge, such (access) rights should be granted within the same timeline, i.e., special access granted within 30 days from generation - being this the general rule of the provision which applies in the context of the exception as well. This is also in line with the rationale of the provision (it would be unreasonable to claim that the "exception" would not be bound to any timeline in the case of an action set up to specifically address a public emergency) and a limitative interpretation of any exception. The use of the exception is also subject to agreement with the funding body and not an elective.

Finally, only in cases where the obligation to protect the result would be impossible (or would require a longer time frame) if the exception is applied, i.e., special access is granted, one could possibly argue that no open access or special access rights would need to be given (or be delayed). However, again in view of the specific context of such grants, any such claim in a concrete case should be carefully examined as it could put in jeopardy the aim of the provision (and possibly put into question by extension the grant). Any abuse should be avoided. In that respect, merely arguing that access rights may endanger protection by a theoretical increased risk (e.g., if a third party would breach its confidentiality requirement) does not seem sufficient (and the beneficiary would clearly not be held responsible under such circumstances).