COVID-RED

innovative medicines initiative

101005177 — COVID-RED

COVID-RED

WP5 – Data management

D5.3 Third iteration of the data management plan

Lead contributor	3 – Julius Clinical
Other contributors	2 - Ava AG

Document History

Version	Date	Description
V1.0	14-05-2021	Deliverable developed for internal review prior to finalization
V2.0	07-06-2021	DMP updated to be in alignment with Consortium Agreement section 7.5.4

COVID-RED

A Data Management Plan created using dmponline

Creators: Billy Franks, Marcel van Willigen

Affiliation: Other

ORCID iD: 0000-0003-4472-4468

Template: Horizon 2020 DMP

Grant number: Proposal Number 101005177

Project abstract:

This project will change the current paradigms combining clinical epidemiologic strategies with digital health approaches in order to detect early symptomatic cases triaging for medical care, efficiently allocating testing capacity and ultimately reducing the time-to-detection of new COVID-19 cases and limiting the risk of disease spread and contamination and improving prognosis for patients. To achieve this, this project presents the first European clinical and digital epidemiology efforts including large cohort studies, digital devices (wearables and mobile app), PCR and antibody testing to allow a fast and reliable detection for COVID-19 carriers and symptomatic individuals suspected of COVID-19 infection.

Last modified: 07-06-2021

COVID-RED - Detailed DMP

1. Data summary

State the purpose of the data collection/generation

Using laboratory-confirmed SARS-CoV-2 infections (detected via serology, PCR and/or antigen tests) as the gold standard, we will determine the sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for each of the following two algorithms to detect *first time* SARS-CoV-2 infection including early or asymptomatic infection: the algorithm using Ava bracelet data when coupled with self-reported Daily Symptom Diary data, and the algorithm using self-reported Daily Symptom Diary data alone. In addition, we will determine which of the two algorithms has superior performance characteristics for detecting SARS-CoV-2 infection including early or asymptomatic infection as confirmed by SARS-CoV-2 virus testing.

Explain the relation to the objectives of the project

The primary endpoint for this study for each subject is the daily indication of potential SARS-CoV-2 infection as provided by the algorithm of the Ava COVID-RED app with or without using data from the Ava bracelet. This daily endpoint will be compared with actual SARS-CoV-2 test results (PCR/antigen and/or serology) collected before, during and at the end of study participation. For the primary comparison, this daily endpoint will be summarized over each trial period per subject to determine (1) whether a subject was ever judged to have had a high risk for a potential SARS-Cov-2 infection, and (2) whether a subject was ever confirmed to have had a SARS-CoV-2 infection by PCR/antigen and/or serology testing. For this comparison, parameters such as sensitivity, specificity, positive predictive value, and negative predictive value will be calculated.

Specify the types and formats of data generated/collected

The primary data types will be:

- Features extracted from raw device data, containing temperature, breathing rate, heart rate variability, pulse rate and/or skin perfusion. (daily)
- Self-reported COVID-19 symptoms. (daily)
- Socio-demographic and basic information of the subject including but not limited to: gender, birth month+year, and household composition. (Once at baseline)
- Electronic CRF (eCRF) data, containing subjects information of length and severity of hospitalization, adverse device effects, and concomitant medication during hospitalization.
- Survey questionnaire data containing information of COVID-19 testing done by participants, COVID-19 behavioral changes, COVID-19 vaccination status, health resource utilization (GP and hospital), COVID-19 related hospitalization events, and bracelet adverse device effects.
- Lab data of Serology and PCR COVID-19 tests taken by participants.

The primary format for most data sources will be comma-separated values (CSV), which is a widely used non-proprietary format. Lab data will be provided in Microsoft excel format. This will ensure long-term usability of the data by us and other parties without the need for specialized software.

Specify if existing data is being re-used (if any)

The algorithms used and further refined in this study have been developed based on proprietary data from the COVI-GAPP pilot study in Liechtenstein (n=1163). Among the 127 seropositive participants at follow-up, sixty-six had worn the Ava bracelet at least 28 days prior to SARS-CoV-2 related symptom onset. An algorithm to detect a SARS-CoV-2 two days prior to symptom onset was trained and validated on this sample, ingesting processed features from the bracelet's raw nightly data. The algorithm had an overall precision of 0.55 and a recall of 0.71 based on the pilot sample.

During the COVID-RED study this algorithm will be further refined for possible detection of SARS-CoV-2 infections, primarily using data collected during the learning phase. Period 1 and Period 2 of the COVID-RED trial will be used to test the algorithm’s performance, where an initial version of the algorithms (v1) will be implemented during Period 1 and an improved version 2 (v2) will be implemented in Period 2. However, data collected during the COVI-GAPP pilot study will not be included in the final COVID-RED FAIR dataset.

Specify the origin of the data

The data will originate from participants who have signed an informed consent to participate in the trial. Data will be collected through self-reporting (Daily Diary in a mobile application and through electronic surveys), by a wearable device that measures physiological parameters, and through self-sampling kits for PCR and serology testing.

Additionally, researchers will contact participants in certain circumstances to conduct a structured interview to support CRF entries on behalf of the participants (e.g., adverse events and hospitalization events) that may contain concomitant medication.

State the expected size of the data (if known)

The final archive is estimated to consist of approximately 100Gb of data.

Outline the data utility: to whom will it be useful

The pseudonymized data will be stored and hosted on the Julius Clinical Data Science Platform (DSP) based on Amazon Web Services (AWS) and the anDREa platform once the clinical trial has been finalized in 2021. The pseudonymized dataset will be a resource for future researchers and trial planners in the area of COVID-19.

Data will become available after initial publications by the COVID-RED consortium (expected Q1-Q2 2022). Research proposals will be evaluated by a committee before access will be granted on the anDREa platform that only contains Pseudonymised data. Approved research proposals will not gain access to personal data or data deemed to be of commercial sensitivity. The anDREa platform also restricts the ability of

sharing/downloading the data unless specifically allowed.

The data will use standards where possible or otherwise contain metadata files to explain the data.

2.1 Making data findable, including provisions for metadata [FAIR data]

Outline the discoverability of data (metadata provision)

The final dataset will be archived by COVID-RED on the Julius Clinical DSP and anDREa platforms. This platform facilitates archiving, online analysis, data sharing, and, in case of anDREa, supports FAIR data principles.

The data will be findable upon publication of the trial results (if not sooner) via use of Open Access publications.

Outline the identifiability of data and refer to standard identification mechanism. Do you make use of persistent and unique identifiers such as Digital Object Identifiers?

Publications resulting from this study in Open Access journals will have a digital object identifier (DOI, https://www.doi.org/).

Outline naming conventions used

Where applicable, standards will be employed such as CDASH. LOINC will be used for lab results if this level of detail is available to the participants. A metadata spreadsheet will be created for data which do not have existing standards (e.g., device extracted features, CRF variables).

Outline the approach towards search keyword

The database references and associated publications will include "COVID-RED" with important keywords such as "COVID-19" to be determined at the time of initial trial publication. A metadata spreadsheet will be supplied to aid identification and discovery of the data.

Outline the approach for clear versioning

Both the anDREa and Julius Clinical DSP support versioning of the data.

Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how

Data will contain extra CSV files referencing the original variable name followed by a description of the variable.

2.2 Making data openly accessible [FAIR data]

Specify which data will be made openly available? If some data is kept closed provide rationale for doing so

Non-pseudonymized data will not be part of the data that will be shared outside the COVID-RED project study team. Only pseudonymized data will be available in the final dataset. The majority of pseudonymized data will be available for approved research proposal which will be evaluated by a committee.

Raw data obtained from the devices and the algorithm will not be made public to protect intellectual properties; instead a subset of features will be extracted and made conditionally available in the final dataset. Data of commercial sensitivity will also be kept closed so as to protectintellectual property and ongoing product development.

We anticipate that the adverse device effects and concomitant medication data will not be made available as this data is collected for regulatory purposes only and not relevant to COVID-19 research.

Specify how the data will be made available

Data will be made available on the Julius Clinical DSP and anDREa platforms and in any publications resulting from this project.

Third parties can get access to the data on the anDREa platform, following proposal review and approval by a committee.

Specify what methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

Both the Julius Clinical DSP and anDREa platforms, include analysis resources and software needed to conduct research in any approved research proposal. The platform includes many common statistical software packages including R and Python, among others.The R analysis environment on the Julius Clinical DSP will be qualified and validated through unit and integration testing.

Specify where the data and associated metadata, documentation and code are deposited

Data and metadata stored on the anDREa platform was setup to be compliant with the FAIR principles. Azure DRE provides Findability (through the Meta Data Catalogue) and Accessibility (through archiving and workflow).
Data and metadata will be stored on the Julius Clinical DSP. The Julius Clinical DSP was set-up with Blob storage ($0.3 per GB per year) that has high durability by design is cost-effective and additional backups are possible, that allows the data to be stored here beyond the lifetime of this grant.

Specify how access will be provided in case there are any restrictions

Data will only be made available for approved research proposals to address the COVID-19 public health emergency and only via the anDREa platform. However, no special access right will be granted to personal data or data of commercial sensitivity. anDREa allows to have separate workspaces for different organisations or research teams. There are costs associated in setting up and maintaining these workspaces.

2.3 Making data interoperable [FAIR data]

Assess the interoperability of your data. Specify what data and metadata vocabularies, standards or methodologies you will follow to facilitate interoperability.

The source data will be in CSV format and include metadata spreadsheets to explain the content and dictionaries in each data table. In some cases, we expect to provide example scripts or programs used for the clinical trial primary analysis to enable future researchers to easily replicate and/or expand published trial results.

Specify whether you will be using standard vocabulary for all data types present in your data set, to allow inter-disciplinary interoperability? If not, will you provide mapping to more commonly used ontologies?

Where applicable, CDASH and LOINC will be used as vocabularies. There will not be a mapping provided to any other ontologies and for data which do not follow CDASH or LOINC there will only be the meta-data spreadsheets available.

2.4 Increase data re-use (through clarifying licenses) [FAIR data]

Specify how the data will be licensed to permit the widest reuse possible

As these data are sourced from participants under informed consent in a healthcare setting, we do not anticipate providing the data in a general fashion under a license agreement, but only after approval of a research proposal by a committee.

Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed

We do plan for data which are specific to an approved research proposal related to the COVID-19 public health emergency to be made available to the proposers of the research. However, all personal data and data of commercial sensitivity will not be made available, regardless of the proposal. Access would be within anDREa platform which means that the data cannot be downloaded or shared further outside of the workspace provided for the researcher(s), unless specifically agreed.

Specify whether the data produced and/or used in the project is useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why

We will not share personal data or data of commercial sensitivity, regardless of the research proposal. We anticipate that initial data sharing would only occur after initial trial publication or aspart of journal review for initial publication.

After research approval, data produced and/or used in this project will be made available to third parties outside the COVID-RED consortium for as long as the anDREa is active and funding is available to store the data there. Per the Consortium Agreement for the COVID-RED trial, only research proposals to address the public health emergency would be eligible for consideration. Furthermore, personal data and data of commercial sensitivity will be not be shared so as to protect participants' identities and COVID-RED researchers' intellectual property.

Describe data quality assurance processes

The data will be reviewed for potential anomalies as part of the trial analysis. Data will be captured via structured forms and guidance will be provided to participants to ensure understanding of the data entry.

There will not be queries to the participant in the event of anomalies or independent monitoring of these data. To prevent data error entries by participants, multiple choices and restricted data entry fields are used whenever possible, while the amount of free-text fields will be reduced to a bare minimum.

eCRF forms for hospitalization events and concomitant medications will be reviewed and queries placed. All concomitant medications will be coded by trained personnel to WHODrug B2 after quality control of the coding takes place by a medical professional. Based on the review of the medical professional the coding will be adjusted until all findings have been resolved.

Publications including data produced in this trial will be peer reviewed by independent reviewers.

Specify the length of time for which the data will remain re-usable

There is no planned date for decommissioning of the trial data. We anticipate that the data will be available for the duration of the anDREa platform's availability.

3. Allocation of resources

Estimate the costs for making your data FAIR. Describe how you intend to cover these costs

Costs have been included within Work Package 5 of the consortium. Data management, technology, and documentation costs will be in excess of 100k euros for this project.

Clearly identify responsibilities for data management in your project

Principle responsibility for data management resides with Julius Clinical Research as part of Work Package 5. Other parties in WP5 contribute to the plans. The device manufacturer and app developer, Ava AG, has a key role in data management related to that subset of the clinical trial data.

Describe costs and potential value of long term preservation

Long-term storage costs are anticipated to be negligible as the anDREa platform utilizes the Microsoft Azure platform where the size of our trial data is "small data" in the realm of cloud storage.

Long-term storage costs are anticipated to be negligible on the Julius Clinical DSP as it utilizes Blob storage ($0.3 per GB per year) and the estimate size of our trial data is considered "small data" in the realm of cloud storage.

Preserving these data for future (independent) research questions is of high scientific and societal value as this research explores SARS-CoV-2 early detection.

4. Data security

Address data recovery as well as secure storage and transfer of sensitive data

All data storages used are protected by firewalls, encryption with access restricted only to

those authorised to process the data. Authorized personnel are identified by the Principle Investigator of the study (Grobbee). Data will not be authorized to be transferred to portable devices (e.g., pendrives) or to personal computers. Instead data transfers will only occur through secured and encrypted methods (e.g., SFTP, HTTPS TSL1.2).

Only pseudonymized data will be loaded into the anDREa and Julius Clinical DSP platforms. No personal identifiers will be available within these data. Data will have beencollected under informed consent which will provide individual approval to use these databeyond a fully anonymized setting in line with GDPR expectations.

The Julius Clinical DSP is a customized platform on Amazon AWS servers based in Frankfurt within the EU. It provides best-in-class security. Both uploading and downloading of data will be under secure protocol. Each upload and downloads of data will be logged (i.e., in an audit trail system is used that can be checked by admins of the Julius Clinical DSP). The Julius Clinical DSP will support Blob storage solution. In addition, data are continuously versioned for at least a 30-day historical period to prevent accidental deletion/overwriting.

The anDREa platform, as a customized platform on MS Azure, provides best-in-class security. Both uploading and downloading of data will be under secure protocol. Each upload and download of data must be approved by the workspace administrators before the system will allow each requested transfer of data (i.e., a ticketing system is used to request specific data transfer actions and only the specific asset can then be transferred upon admin approval).

The anDREa platform, as a customized platform on MS Azure, provides best-in-class backup and availability of the data. In addition, data are continuously versioned for a 30-day period. We will use a "source storage" and "working storage" approach to ensure that the data being used in the "working storage" are not incidentally (and unknowingly) changed in comparison to the "source storage" area.

The key to link between the unique participant identifier and the database containing identified participant data will not be stored or shared outside of Julius Clinical Research (the site for the purposes of this decentralized trial). In addition, the link will not be stored in the Julius Clinical DSP nor in any other systems where the pseudonymized trial data will be analyzed. The link will be maintained only in those systems within Julius Clinical Research where the identified participant data are stored. These internal systems will be accessible only by Julius Clinical Research staff who have permission to view the identified participant data for the purpose of distribution of clinical trial materials and follow-up with participants for any reported adverse event or hospitalization events.

The Ava COVID-RED app was specifically designed for the COVID-RED clinical trial. Its underlying codebase and backend architecture, however, is largely based on the Ava Fertility Tracker app, which, together with the Ava bracelet, comprises the Ava Fertility Tracker medical device. Both Ava COVID-RED and the Ava Fertility Tracker apps are manufactured by Ava AG, Gutstrasse 73, 8055, Zurich, Switzerland. App-recorded data will be stored on a European instance of Amazon Web Services (AWS) in Frankfurt, Germany. We intend on exporting from these servers to Andrea a CSV file with de-identified, aggregated data points (at up to 1 reading of each physiological parameter of interest per participant per day, plus any daily diary entries) at pre-designated time points.

5. Ethical aspects

To be covered in the context of the ethics review, ethics section of DoA and ethics deliverables. Include references and related technical aspects if not covered by the former

Only pseudonymized data will be loaded into the anDREa and Julius Clinical DSP. No personal identifiers will be available within these data. Data will have been collected under informed consent which will provide individual approval to use these data beyond a fully anonymized setting in line with GDPR expectations. The informed consent will include provisions for the long-term storage by the site and future research use of their pseudonymized data.

6. Other

Refer to other national/funder/sectorial/departmental procedures for data management that you are using (if any)

Data collection and data management will be conducted under the applicable SOPs of Julius Clinical Research and Ava AG, depending on the data source. Within the consortium, data sharing procedures and agreements will be installed to ensure compliance with GDPR. COVID-RED will take part in the Open Research Data Pilot. As such, we will ensure that data generated in this project will be FAIR and accessible with as few restrictions as possible, while at the same time protecting sensitive data from inappropriate access.