COVID-RED

101005177 — COVID-RED

COVID-RED

WP5 – Data management

D5.1 First iteration of the data management plan

Lead contributor	3 – Julius Clinical
Other contributors	1 – UMCU 2 – Ava AG

Document History

Version	**Date**	**Description**
V1.0	15-09-2020	Deliverable description developed for internal review prior to finalization

Contents

1. Data summary

Provide a summary of the data addressing the following issues:

- **State the purpose of the data collection/generation**
- **Explain the relation to the objectives of the project**
- **Specify the types and formats of data generated/collected**
- **Specify if existing data is being re-used (if any)**
- **Specify the origin of the data**
- **State the expected size of the data (if known)**
- **Outline the data utility: to whom will it be useful**

- The purpose of data collection is to evaluate the performance of the device and digital apps that are used within this clinical trial.
- The database will be the basis for the conduct of statistical testing and statistical analysis of the clinical trial.
- The primary data type will be features extracted from raw device data, symptoms and demographic+medical history data reported into an app, health resource utilization data, electronic CRF data, lab results data, survey data, and adverse events data. The primary format for the data sources will be CSV.
- The database may incorporate data from the Corona Check App (https://decoronacheck.nl/).
- The data will originate directly from participants who have completed an informed consent to participate in the trial. Additionally, researches will contact participants in certain circumstances to conduct a structure interview to support CRF entries on behalf of the participants (e.g. adverse events and hospitalization events).
- The final archive is estimated to consist of approximately 100Gb of data.
- The dataset will provide a resource for future researchers and trial planners in the area of COVID-19.

2. FAIR data

2.1 Making data findable, including provisions for metadata:

- **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?**
- **Outline naming conventions used**
- **Outline the approach towards search keyword**
- **Outline the approach for clear versioning**
- **Specify standards for metadata creation (if any). If there are no standards in your discipline describe what metadata will be created and how**

- The final dataset will be archived by COVID-RED on the anDREa platform (https://www.andrea-consortium.org/about-andrea/). This platform facilitates archiving, online analysis, data sharing, and supporting FAIR data principles.
- The data will be findable upon publication of the trial results (if not sooner) by use of the anDREa Meta Data Catalogue and via use of Open Access publications.
- 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.
- Where applicable CDASH standards will be employed (e.g. CRF elements) and LOINC 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).

2.2 Making data openly accessible:

- **Specify which data will be made openly available? If some data is kept closed provide rationale for doing so**
- **Specify how the data will be made available**
- **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)?**
- **Specify where the data and associated metadata, documentation and code are deposited**
- **Specify how access will be provided in case there are any restrictions**

- The majority of data will be available for approved research proposal which will be evaluated by a committee.
- Of the data listed, we anticipate that the adverse events data will not be made available as this data is collected for regulatory purposes only and not relevent to COVID-19 research.
- The anDREa platform includes analysis resources and software needed to conduct research in any approved research proposal. The platform includes many common statistical software packages including R, Python, and julia.
- Data and metadata will reside in the same platform as the analysis platform (anDREa).
- Data will only be made available for approved research proposals and only via the anDREa platform.

2.3 Making data interoperable:

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

- 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 programs which were used for the clincial trial primary analysis to enable future researchers to easily replicate and/or expand published trial results.
- 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):

- **Specify how the data will be licenced to permit the widest reuse possible**
- **Specify when the data will be made available for re-use. If applicable, specify why and for what period a data embargo is needed**
- **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**
- **Describe data quality assurance processes**
- **Specify the length of time for which the data will remain re-usable**

- As these data are sourced from partipants under informed consent in a healthcare setting, we do not anticipate providing the data in a general fashion under a license agreement.
- We do plan for data which are specific to any given research proposal to be made available to the proposers of the research. Access would be within anDREa which means that the data cannot be downloaded or shared further outside of the workspace provided for the researcher(s).

- We anticipate that initial data sharing would only occur after initial trial publication or as part of journal review for initial publication.
- 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.
- 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

Explain the allocation of resources, addressing the following issues:

- **Estimate the costs for making your data FAIR. Describe how you intend to cover these costs**
- **Clearly identify responsibilities for data management in your project**
- **Describe costs and potential value of long term preservation**

- 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.
- Principle responsibility for data management resides with Julius Clinical Research as part of Work Package 5. Other parties in WP5 contribute to the plans and the device manufacturer and app developer, Ava, have a key role in data management related to that subset of the clinical trial data.
- 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. Preserving these data for future (independent) research questions is of high scientific and societal value as this research explores COVID-19 early detection.

4. Data security

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

- Only pseudo-anonymized data will be loaded into the anDREa platform. 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 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 downloadsof 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 addtion, 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.

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 pseudo-anonymized data will be loaded into the anDREa platform. 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 and future research use of their 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. 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.

History of change

IMI2/INT/2016-00954	Version dated 2016
IMI2/INT/2016-00954 v. 2019:	Updated version Simplification cover page Replace ''publishable summary'' by ''summary.to avoid any confusion with the dissemination level of the deliverable.