ORCHESTRA

Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic

WP10_D10.4,

Anonymous dataset for the public published

UHC

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

Project Classification

Project Acronym:	ORCHESTRA
Project Title:	Connecting European Cohorts to Increase Common and Effective Response to SARS- CoV-2 Pandemic
Coordinator:	UNIVR
Grant Agreement Number:	101016167
Funding Scheme:	Horizon 2020
Start:	1st December 2020
Duration:	36 months
Website:	www.orchestra-cohort.eu
Email:	info@orchestra.eu

Document Classification

WP No:	WP10
Deliverable No:	D10.4
Title:	Anonymous dataset for the public published
Lead Beneficiary:	UHC
Other Involved Beneficiaries:	UNIVR, SAS, CINECA, RER-ASSR, ISGLOBAL, LMU MUENCHEN, UANTWERPEN, USTUTT, CERMEL, UMCG, CINES, FCRM
Nature:	Publication
Dissemination Level:	Public
Due Delivery Date:	31 January 2022
Submission Date:	20 May 2022
Justification of delay:	Dependency on central data availability
Status:	Completed
Version:	1.0
Author(s):	Sina Hopff, Janne Vehreschild, Chin Lee, Carolin Jakob

History of Changes

Version	Date	Created/Modified by
0.1	29 April 2022	Sina Hopff
0.2	04 May 2022	Fabian Prasser
1.0	20 May 2022	Sina Hopff, Carolin Jakob

Table of contents

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

Executive Summary

The key task of WP10 is dissemination of scientific content to various stake holders, with the goal to attract ideas and collaboration from the scientific community and inform decision makers on need for action. This task has been conducted jointly with UNIVR and all other ORCHESTRA partners.

Between October 2021 and April 2022, the ORCHESTRA anonymization pipeline was established. After setting up the data protection concept, the documents were submitted to a law firm for a legal advice at 24th January 2022. The delay of the task resulted from the late availability of central data. The first central dataset was sent to UHC on 10th February 2022. Therefore, the previously devised anonymization steps could be applied to the raw data starting in February 2022 to create a first anonymous dataset that will be publicly available on the ORCHESTRA website. More complex datasets are now created analogously to the ORCHESTRA data protection concept.

We received the legal advice of the ORCHESTRA Data Protection Concept on 18th May 2022. The summary of the results of the legal opinion states the following: “The data protection concept ("General Data Protection Concept for Legal Opinion") largely complies with the legal requirements for the processing of special categories of personal data in the form of health data in connection with a clinical study. However, there is still a need to specify individual points.” The proposals for concretization will now be incorporated in detail into the data protection concept. This requires detailed agreements between the persons involved and thus a certain amount of time, so that the final implementations will be submitted later.

A preliminary version of the PUF is attached in Appendix 1 (ORCHESTRA-raw-dataset_2022-05-20.xls) and Appendix 2 (ORCHESTRA-puf-dataset_2022-05-20.xls). The corresponding analysis for the ORCHESTRA website can be accessed via the following link. The data are not yet published at the current time.

ORCHESTRA PUF analysis

General Data Protection Concept for Legal Opinion

1. General Information about ORCHESTRA

1.1. Introduction ORCHESTRA-Cohort

ORCHESTRA is a three-year international research project aimed at tackling the coronavirus pandemic, led by the University of Verona and involving 26 partners (extending to a wider network of 37 partners) from 15 countries: Argentina, Belgium, Brazil, Congo, France, Gabon, Germany, India, Italy, Luxemburg, Netherlands, Romania, Slovakia, Spain, Venezuela. The

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

project is funded by the European Union’s Horizon 2020 research and innovation program under the ERAvsCORONA Action Plan which was developed jointly by Commission services and national authorities. The main outcome of ORCHESTRA is the creation of a new pan-European cohort built on existing and new large-scale population cohorts in European and non-European countries. Data analysis through a federated learning technique supported by advanced modelling capabilities will allow the integration of epidemiological, clinical, microbiological and genotypic aspects of population-based cohorts with environment and socio-economic features. The ORCHESTRA cohort will include SARS-CoV-2 infected and non-infected individuals of all ages and conditions and thereby enabling a retrospective evaluation of risk factors for the disease acquisition and progression of the disease and prospective follow-up aimed at exploring long term consequences and analysis of vaccination response. To better address these research questions, the ORCHESTRA-Cohort will include adequately sampled representatives of general populations, COVID-19 patients and special ‘at risk’ populations of fragile individuals and health-care workers [1].

1.2. European guidance framework for clinical decision-making algorithms in ORCHESTRA

- Council of Europe, Recommendation CM/Rec (2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers’ Deputies) [2]
- EU, High-Level Expert Group on Artificial Intelligence, ETHICS GUIDELINES FOR TRUSTWORTHY AI, 2019 [3]
- COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Building Trust in Human-Centric Artificial Intelligence, Brussels, 8.4.2019 COM (2019) 168 final [4]
- ICDPPC, DECLARATION ON ETHICS AND DATA PROTECTION IN ARTIFICIAL INTELLIGENCE, 40th International Conference of Data Protection and Privacy Commissioners, Tuesday 23rd October 2018, Brussels [5]
- EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES, Statement on European Solidarity and the Protection of Fundamental Rights in the COVID-19 Pandemic (2/4/2020) [6]

1.3. Data sharing

As within ORCHESTRA a large amount of data will be handled, a dedicated Work Package has been designed for data management and data protection. The Technology Partners of the project will act as major actors. Data collection, production and archiving will assure alignment to FAIR principles and ensure that FAIR data practices are applied and maintained through the project and in the post project sustainability planning. The Data Management Plan team will also ensure that the project is managed in accordance with the H2020 open access publication strategy and the EU Public Health Emergency strategy for data sharing [1].

1.4. Data generation and collection

ORCHESTRA will use available datasets but also data collected during the project and will analyze secondary data that are publicly available or for which agreements will be established during the project. Data will be formalized in structured databases for the purposes of elaborations to be carried out in the project. They will be used for achieving project objectives and related scientific dissemination activities, including follow up scientific activities.

1.5. GECCO Dataset

The GECCO (German Corona Consensus Data Set) research dataset on COVID-19 developed by the German COVID-19 research network "Netzwerk Universitätsmedizin" has been selected as the basis for creating a core dataset in ORCHESTRA. The German Corona Consensus Data Set (GECCO) is a German uniform dataset for the systematic collection of scientific data on COVID-19 in Germany. Basis of the dataset are both the ISARIC protocol of the WHO [7] as well as the data concepts of the Lean European Open Survey on SARS-CoV-2 infected patients (LEOSS) registry [8]. The purpose of ISARIC is to prevent disease and death from outbreaks of infectious diseases such as COVID-19. ISARIC brings together clinical research networks worldwide to provide the fastest possible research response to an outbreak of an infectious disease. To ensure syntactic and semantic interoperability, the data elements are mapped to international standards and terminology [9].

1.6. Strategy for data management and protection

Data management plays a major role in the project. Data Cohorts in the projects will be available in a wide variety of platforms and data models. In this regard, data harmonization, normalization and consolidation appropriate and efficient workflow is of paramount importance. To facilitate these efforts, a central data repository will be set up for sharing, aggregating, integrating and analyzing data. Each National Data Provider will be responsible for data production and anonymization and pseudonymization of all the sensitive data as required by national laws. According to the information available within the Consortium, Article 26 of the GDPR (Joint controllership) is not applicable. Every Data Provider will be autonomous data controller of their own data and each National Hub Providers will be nominated data processor by their own corresponding Data Providers, according to Article 28 of the GDPR. ORCHESTRA recognizes the value of regulating research data management issues and foresees the delivery of a Data Management Plan (DMP). Every partner will be responsible for the secure storage of the data he generates in reliable facilities that permit recovery and do not compromise their integrity. The ORCHESTRA data architecture, as the integration of the cloud-based infrastructure developed in WP7 will be compliant by design with the highest standards for privacy data protection. For the anonymized dataset (Public Use File (PUF) and Scientific Use

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

File (SUF)), comprehensive measures will be taken to rule out risk of re-identification. Such measures could have the following framework [1].

1.7. Properties of the data

The Data is either transferred based on informed consent (pseudonymized) or in an anonymous form. Clinical data are (i) from clinical routine or (ii) as part of prospective epidemiological studies. Source data represent patient characteristics, interventions, medications, clinical course of a disease, and outcome of patients with COVID-19 infections and/or defined underlying disease and/or health workers. None of the participating cohorts will cover 100% of all members of the represented cohort, but can be considered a random sample from a much larger overall group. For all cohorts, it can be presumed that more than 100 other cases fulfilling the inclusion criteria exist in the general population for one cohort member.

2. Implementation concept of the Public and Scientific Use Files

2.1. Project description and aims of the provision of the data

The first purpose of the project is to provide a publicly available narrow dataset (PUF). The dataset represents the clinical development of defined patient populations and thus enables predictions, feasibility analyses and clinical trial design. The dataset will be available for download and online analysis via a dashboard. The risk of re-identification is sufficiently reduced by an anonymization pipeline. The PUF should be downloadable with the CC4.0 license.

Furthermore, comprehensive scientific datasets (SUF) will be generated. They are tailored to specific research questions. Accredited and established research groups can gain access to such datasets for a given challenge after signing a license agreement. Participants of the challenges will be connected via different communication channels. Results can be presented to the ORCHESTRA study group and stake holders during science slams and winners may be elected to publish the official ORCHESTRA solution to a given problem. By using a common basic dataset, the SUF can be transferred to other concepts in ORCHESTRA. The risk of re-identification is reduced by an anonymization pipeline and additionally the risk of misuse is reduced by a data use and access statement. Access to the data is controlled and shared with researchers within Europe after justification of the research project.

2.2. Guidelines for anonymization

- NHS Anonymization Standard for Public Health and Social Care Data [10]
- Opinion 05/2014 on Anonymization Techniques by the Article 29 Data Protection Working Party [11]
- European Medicines Agency's Policy 007 Implementation Guideline [12]

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

2.3. Anonymization pipeline

The anonymization pipeline is used to create the PUF and SUF. The following steps describe the anonymization process:

Step 1: Exclusion of any directly identifying information

1. The dataset will not contain any directly identifying information, i.e. no name, birth date, place of living, insurance company or any identifiers linked to the patient record or any other file at the participating center.
2. It will not contain any specific dates. All major events will be recorded as days from diagnosis, almost all data will be aggregated over major stages of the infection.
3. The start date of the observational period, such as the day of diagnosis, hospitalization or treatment, is recorded in periods, e.g. quartiles (PUF) or months (SUF).
4. No free text items will be included.
5. No information about the documented/treated site is included, only country and type of hospital. Each PUF/SUF contains clinical data from a minimum three sites.

Step 2: Categorization

1. All metric data will be rounded and/or summarized into value groups to limit attacks using reverse database search in hospital information systems
2. Dichotomous data associated with considerable risk of re-identification, such as specific diagnosis, will be grouped into, for example, cardiovascular diseases (see table "PUF-Table - Assessment of the re-identification risk associated with individual variables_2022-01-20") or more granular answers depending on the specific research question (see table "SUF-Table - Assessment of the re-identification risk associated with individual variables_2022-01-20").

Step 3: Sensitive data with risk for discrimination

1. Sensitive data that can lead to discrimination of individuals, among other things, will be protected from inference, e.g. human immunodeficiency virus (HIV) or genetic diseases.

Step 4: Qualitative risk assessment

1. Variables will be assessed with respect to their replicability, availability, and distinguishability to obtain their privacy risk (quantified by 1=low, 2=medium, 3=high) according to Malin et al (2011) [13].
2. Variables will be evaluated according to their risk of inference.

Step 5: Quantitative risk assessment

1. Variables with a sum of weight > 6 in replicability, availability and distinguishability will be defined as key variables. The key variables are protected from attacks using singling

out and linkage using k-anonymity (k=11 according to the Opinion 05/2014 on Anonymization Techniques by the Article 29 Data Protection Working Party and the European Medicines Agency's Policy 007 Implementation Guideline) [11, 12].

2. Variables which were defined as being associated with a risk of inference are protected using t-closeness. We selected t=0.5 according to Jakob et al. (2020) [14].
3. Variables which are perfectly correlated to other variables do not need additional protection from attacks using inference. Parameters which are collected at different time points, such as symptoms at diagnosis and at discharge are considered as highly correlated. Reason is that 1) for dichotomous parameter in most cases the information that a parameter was not present is no information which increase the risk for re-identification. 2) If a parameter was present the probability increases that the parameter was present also at a later time point. 3) For metric parameter, the probability is high that the value at a later time point is influenced by the values at an earlier time point.

2.4. Description k-Anonymity

K-Anonymity ensures that each record is indistinguishable from at least k-1 other records regarding the key variables, i.e. variables that could be used for linkage [15]. The Article 29 Data Protection Working Party recommends a value of k > 10, which is consistent with recommendations from other guidelines, including the European Medicines Agency's Policy 007 Implementation Guideline [12], which recommends a risk threshold of 0.09 (corresponding to k=11).

2.5. Description t-Closeness

For the variables that need to be protected from inference, we implement the well-known t-closeness model [16] with t=0.5. This approach has been recommended by the opinion [2] and the parameterisation considers the high level of privacy protection already achieved. By combining protection against singling out and linkage with additional protection against inference of sensitive information, the resulting dataset is strongly protected from the threats addressed by relevant guidelines and laws.

2.6. Justifications that the risk of re-identification is sufficiently controlled with the described anonymization concept

To make sure that the risk of re-identification is controlled, quantitative and mathematically comprehensible calculation of the risk of re-identification are made. The background population of the considered patient data is huge as the consortium covers Europe and non-EU countries. In the WP2 long-term sequelae around 10.000 patients will be included. The current number of cases is 5976 (as of 26th October 2021). The pipelines are oriented towards current guidelines [10-12]. Furthermore, in the COVID-19 pandemic, the anonymization concept was accepted within the Lean European Open Survey on SARS-CoV-2 infected

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

patients (LEOSS) and applied to clinical data from over 130 European sites [8]. The concept will also be implemented for the German National Pandemic Cohort Network (NAPKON).

2.7. Variables PUF

- Baseline
    - Quarter of diagnosis
    - Demographics: age, gender, BMI
    - Comorbidities: pulmonal, cardial, nephrological, hemato-oncological, rheumatological/immunological, liver diseases, diabetes mellitus, transplantation indicated in groups (yes/no/unknown)
    - Smoking ever (yes/no/unknown)
    - SARS-CoV-2-vaccination (yes/no/unknown)
- Acute phase (all items are considered in relation to the acute phase)
    - COVID-19-specific therapies (yes/no/unknown)
    - Dialysis (yes/no/unknown)
    - Intensive care treatment (yes/no/unknown)
    - Venous thrombosis (yes/no/unknown)
    - Pulmonary embolism (yes/no/unknown)
    - Stroke (yes/no/unknown)
    - Myocardial infarction (yes/no/unknown)
    - Pulmonary co-infection (yes/no/unknown)
    - Stage at diagnosis (regarding to WHO progression scale, mild, moderate, severe, death)
    - Most severe stage (regarding to WHO progression scale, mild, moderate, severe, death)
    - Time diagnosis until hospitalization (days)
    - Time diagnosis until stage moderate (days; first time fulfilling criteria)
    - Time diagnosis until stage severe (days; first time fulfilling criteria)
    - Time diagnosis until invasive ventilation (days; first time fulfilling criteria)
    - Total days hospitalized (days)
    - Total days invasive ventilation (days)
    - Any symptom (yes/no/unknown)
    - Symptoms in acute phase (D0, D7, D14) and at the end of acute phase(general, neurological, respiratory, gastrointestinal) (yes/no/unknown)
    - Vital signs in acute phase (D0, D7, D14) and at the end of acute phase (blood pressure syst/diast, heart frequency, peripheral oxygen saturation, respiratory frequency in categories, fever yes/no)
- Endpoints follow-up visits (visit time calculated from date of initial diagnosis): after 3, 6 and 12 months
    - Asymptomatic (yes/no)
    - Symptoms (general, neurological, respiratory, gastrointestinal) (yes/no)

ORCHESTRA has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016167

- Vital signs (blood pressure syst/diast, heart frequency, peripheral oxygen saturation, respiratory frequency in categories, fever yes/no)
- Type of discharge (alive / admission to hospital / referral to another institution / death / unknown)

2.8. Dataset for SUF

- could include all visit points: Baseline, acute phase, end of acute phase, all follow-up visits
- could include all items of the GECCO83 datasets according to table “SUF-Table - Assessment of the re-identification risk associated with individual variables_2022-01-20” plus items to classify the severity of the disease according to the WHO progression scale

References

1. CORDIS EU research results - ORCHESTRA. Available from: https://cordis.europa.eu/project/id/101016167.

2. Council of Europe, Recommendation CM/Rec (2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems (Adopted by the Committee of Ministers on 8 April 2020 at the 1373rd meeting of the Ministers’ Deputies) Available from: https://rm.coe.int/09000016809e1154.

3. EU, High-Level Expert Group on Artificial Intelligence, ETHICS GUIDELINES FOR TRUSTWORTHY AI, 2019 Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

4. COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS Building Trust in Human-Centric Artificial Intelligence, Brussels, 8.4.2019 COM (2019) 168 final Available from: https://ec.europa.eu/jrc/communities/sites/jrccties/files/ec_ai_ethics_communication_8_april_2019.pdf.

5. ICDPPC, DECLARATION ON ETHICS AND DATA PROTECTION IN ARTIFICIAL INTELLIGENCE, 40th International Conference of Data Protection and Privacy Commissioners, Tuesday 23rd October 2018, Brussels Available from: https://privacyconference2018.org/system/files/2018-10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf.

6. EUROPEAN GROUP ON ETHICS IN SCIENCE AND NEW TECHNOLOGIES, Statement on European Solidarity and the Protection of Fundamental Rights in the COVID-19

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167

Pandemic (2/4/2020) Available from: https://ec.europa.eu/info/research-and-innovation/strategy/support-policy-making/scientific-support-eu-policies/ege_en.

7. ISARIC protocol. Available from: https://isaric.net/ccp/.

8. LEOSS-Registry Data Protection Concept. Available from: https://leoss.net/wp-content/uploads/2020/04/LEOSS-Public-Use-File-Protection-2020-04-21-1.pdf.

9. Article 29 working party archives 1997 - 2016. Available from: https://ec.europa.eu/justice/article-29/documentation/index_en.htm.

10. NHS Anonymisation Standard for Publishing Health and Social Care Data. Available from: https://digital.nhs.uk/data-and-information/information-standards/information-standards-and-data-collections-including-extractions/publications-and-notifications/standards-and-collections/isb1523-anonymisation-standard-for-publishing-health-and-social-care-data.

11. ARTICLE 29 DATA PROTECTION WORKING PARTY. 0829/14/EN WP216. Opinion 05/2014 on Anonymisation Techniques. 2021-11-23]; Available from: https://www.pdpjournals.com/docs/88197.pdf.

12. External guidance on the implementation of the European Medicines Agency policy on the publication of clinical data for medicinal products for human use. EMA/90915/2016 Version 1.4. 2018 Available from: https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data.

13. Malin, B., et al., Identifiability in biobanks: models, measures, and mitigation strategies. Hum Genet, 2011. 130(3): p. 383-92.

14. Jakob, C.E.M., et al., *Design and evaluation of a data anonymization pipeline to promote Open Science on COVID-19.* Sci Data, 2020. **7**(1): p. 435.

15. Sweeney, L., k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 557-570.

16. Li, N., T. Li, and S. Venkatasubramanian, t-closeness: Privacy beyond k-anonymity and l-diversity. 23rd International Conference on Data Engineering, 2007: p. 106-115.

Acknowledgments

Elisa Gentilotti, Evelina Tacconelli, UNIVR

Chin Lee, Nick Schulze, Marlene Nunnendorf, Carolin Jakob, Luise Wolf, Janne Vehreschild, UHC

Fabian Prasser, BIH at Charité Berlin

ORCHESTRA has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101016167