

International Clinical Research Data Ecosystem: From Data Standardization to Federated Analysis

Eugenia RINALDI^a, Chiara DELLACASA^b, Miroslav PUSKARIC^c, Thomas OSOMO^d,
Anna GORSKA^e and Caroline STELLMACH^a

^aBerlin Institute of Health at Charité-Universitaetsmedizin Berlin, Germany

^bCineca Consorzio Interuniversitario, Bologna, Italy

^cHigh-Performance Computing Center Stuttgart, Germany

^dCentre Informatique National de l'Enseignement Supérieur, Montpellier, France

^eUniversity of Verona, Italy

ORCID ID: Eugenia Rinaldi <https://orcid.org/0000-0003-0343-6400>

Abstract. Within the HORIZON 2020 project ORCHESTRA, patient data from numerous clinical studies in Europe related to COVID-19 were harmonized to create new knowledge on the disease. In this article, we describe the ecosystem that was established for the management of data collected and contributed by project partners. Study protocols elements were mapped to interoperability standards to establish a common terminology. That served as the basis of identifying common concepts used across several studies. Harmonized data were used to perform analysis directly on a central database and also through federated analysis when data was not permitted to leave the local server(s). This ecosystem facilitates the answering of research questions and generation of new knowledge available for the scientific community.

Keywords. Clinical research, FAIR data, Interoperability, Federated data analysis, Data management, Data transformation

1. Introduction

The project ORCHESTRA [1] aims to generate new knowledge on COVID-19 by creating a pan-European knowledge base derived data collected in numerous European clinical studies. To do so, an elaborate data infrastructure needed to be established to enable: data collection, data harmonization and standardization [2], as well as data storage and analysis, and last data protection.

2. Methods

The REDCap platform was selected for data collection and storage. We mapped the data elements across the different ORCHESTRA studies to international terminology standards (e.g. SNOMED CT [3], LOINC [4]) and added the corresponding standard codes to the metadata.. To investigate a specific research question, all the data collected using relevant variables were transformed to a common format and merged. REDCap

provides direct access for centralized analysis. Additionally, a pipeline for federated data analysis [5] was built, using DataShield libraries for privacy preserving analysis of sensitive data.

3. Results

To enable joint analysis across studies, a harmonization process of the clinical study variables that comprise the data collection forms was set up. Semantic interoperability [6,7] in ORCHESTRA studies was established based on standard terminology. This facilitated the next step of harmonization and identification of similar concepts across studies that were candidates for shared analysis.

4. Discussion

Considering the importance of cross-countries studies to face infectious diseases, it would be very useful to reach international agreements on standard formats for common elements used in the clinical studies on infectious diseases.

5. Conclusions

Harmonized datasets could be used to gain new knowledge on the disease, both performing analysis on a central server and using federated approach. The use of standards supports the joint analysis across studies and facilitates the use of analysis tools and federated analysis algorithms.

References

- [1] Connecting European Cohorts to Increase Common and Effective Response to SARS-CoV-2 Pandemic: ORCHESTRA | ORCHESTRA Project | Fact Sheet | H2020 | CORDIS | European Commission [Internet]. [cited 2022 Mar 14]. Available from: <https://cordis.europa.eu/project/id/101016167>
- [2] Rinaldi E, Stellmach C, Rajkumar NMR, Carocchia N, Dellacasa C, Giannella M, et al. Harmonization and standardization of data for a pan-European cohort on SARS- CoV-2 pandemic. *NPJ Digit Med*. 2022 Jun 14;5(1):75.
- [3] SNOMED CT: working smarter, not harder - PubMed [Internet]. [cited 2023 Aug 21]. Available from: <https://pubmed.ncbi.nlm.nih.gov/36702604/>
- [4] Rychert J. In support of interoperability: A laboratory perspective. *Int J Lab Hematol*. 2023 Aug;45(4):436–41.
- [5] Narmadha K, Varalakshmi P. Federated Learning in Healthcare: A Privacy Preserving Approach. *Stud Health Technol Inform*. 2022 May 25;294:194–8.
- [6] Gansel X, Mary M, van Belkum A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. *Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol*. 2019 Jun;38(6):1023–34.
- [7] Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. *Npj Digit Med*. 2019 Aug 20;2(1):1–5.