Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response

Evelina Tacconelli,[a] Anna Gorska,[a] Elena Carrara,[a] Ruth Joanna Davis,[a] Marc Bonten,[b] Alex W. Friedrich,[c,d] Corinna Glasner,[c] Herman Goossens,[e] Jan Hasenauer,[f,g] Josep Maria Haro Abad,[h,i] José L. Peñalvo,[j] Albert Sanchez-Niubo,[h,k] Anastassja Sialm,[l] Gabriella Scipione,[m] Gloria Soriano,[j] Yazdan Yazdanpanah,[n] Ellen Vorstenbosch,[h,i] and Thomas Jaenisch [o]*

[a]Infectious Diseases Division, Department of Diagnostics and Public Health, University of Verona, Verona, Italy
[b]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Netherlands
[c]Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands
[d]University Hospital Münster, Münster, Germany
[e]Laboratory of Medical Microbiology, Vaccine and Infectious Diseases Institute, University of Antwerp, Antwerp, Belgium
[f]Life and Medical Sciences Institute, University of Bonn, Bonn, Germany
[g]Institute of Computational Biology, Helmholtz Center Munich - German Research Center for Environmental Health, Neuherberg, Germany
[h]CIBER en Salud Mental (CIBERSAM). Instituto de Salud Carlos III, Madrid, Spain
[i]Parc Sanitari Sant Joan de Déu, Sant Boi de Llobregat, Spain
[j]Department of Public Health, Institute of Tropical Medicine, Antwerp, Belgium
[k]Department of Social Psychology and Quantitative Psychology. Faculty of Psychology, University of Barcelona, Barcelona, Spain
[l]Schweizer Paraplegiker Forschung, Nottwil, Switzerland
[m]Supercomputing Applications and Innovation Department, Cineca Consorzio Interuniversitario, 40033 Casalecchio di Reno, Italy
[n]INSERM, IAME, Hôpital Bichat - Claude-Bernard, Infectious Diseases Department, France
[o]Heidelberg Institute of Global Health, Heidelberg University Hospital, Heidelberg, Germany

Summary

The COVID-19 pandemic saw a massive investment into collaborative research projects with a focus on producing data to support public health decisions. We relay our direct experience of four projects funded under the Horizon2020 programme, namely ReCoDID, ORCHESTRA, unCoVer and SYNCHROS. The projects provide insight into the complexities of sharing patient level data from observational cohorts. We focus on compliance with the General Data Protection Regulation (GDPR) and ethics approvals when sharing data across national borders. We discuss procedures for data mapping; submission of new international codes to standards organisation; federated approach; and centralised data curation. Finally, we put forward recommendations for the development of guidelines for the application of GDPR in case of major public health threats; mandatory standards for data collection in funding frameworks; training and capacity building for data owners; cataloguing of international use of metadata standards; and dedicated funding for identified critical areas.

The Lancet Regional Health - Europe 2022;21: 100467
Published online 4 August 2022
https://doi.org/10.1016/j.lanepe.2022.100467

Keywords: **Data sharing; General Data Protection Regulation; SARS-CoV-2; Cohort study; Preparedeness; Pandemic; Machine learning**

Introduction

The COVID-19 pandemic has underlined the importance of collaborative efforts to effectively address new public health threats. At the same time, it has highlighted major stumbling blocks to such collaborative efforts in terms of access to and interoperability of heterogeneous health-related data.

*Corresponding author at: University of Verona: Universita degli Studi di Verona, Diagnostics and Public Health, Piazzale Ludovico Antonio Scuro n.10, 37134 Verona, Italy.
E-mail address: evelina.tacconelli@univr.it (E. Tacconelli).

In this viewpoint article, we provide an overview of the challenges encountered when sharing health data within collaborative research projects, looking at participant level clinical data from observational research studies as well as at patient-level data from Electronic Health Records (EHR). Against the backdrop of the research funding response to the COVID-19 pandemic on the part of the European Commission (EC), we highlight the potential that such shared data offers in terms of enhancing our knowledge about the pandemic, by introducing advanced methodologies and flexibility in the analyses. At the same time, we consider the constraints of sharing and utilizing these data which is identified as highly sensitive and carries extra levels of protection.

We discuss these challenges from the perspective of the investigators of a selection of real-life projects implemented prior to and during the pandemic. We describe a number of solutions that have been employed in these projects to partially overcome data sharing challenges, such as federated data analysis (or federated learning) and the adoption of common standards for data encoding and sharing. In addition, we assess how research funding frameworks and conditions could be adapted to facilitate data sharing with a view to a faster delivery of public-health relevant results. Finally, we develop a list of concrete action points to enhance data sharing processes to prepare for and respond to the current and future public health emergencies.

The European COVID-19 research investment landscape

As the reality of the novel coronavirus SARS CoV-2 outbreak began to hit home at the dawn of 2020, the European Commission (EC) and national governments alike, were united in their commitment to invest in coordinated scientific research as a first weapon to tackle the new disease. Already by March 2020, emergency research funding of €48.5m for a total of 18 projects was made available through a fast-track call for expression of interest[1] aimed at understanding the virus, better clinical management of patients and improved public health preparedness. In parallel, the European Innovative Medicines Initiative (IMI) launched a special fast-track call for the “Development of therapeutics and diagnostics combatting coronavirus infections” with a budget currently amounting to €117m. As the epidemiological situation evolved, so did the direction and focus of research. By April 2020, the first ERAvsCORONA short term action plan[2] was born, pledging to prioritise areas such as EU-wide clinical trials, data sharing and global collaboration. In the frame of this action plan, a second call for expression of interest[3] was launched in May 2020 with a budget of €129.5m and led to the financing of an additional 23 projects. However, the projects awarded under these calls (e.g. RECOVER,[4] ORCHESTRA,[5] UNCOVER,[6] EU-RESPONSE[7] to name a few) were only just taking off when the epidemiological situation once again began to rapidly evolve with the launch of the vaccination campaigns and the emergence of multiple SARS-CoV-2 variants. In response, the EC launched the European Health Emergency Preparedness and Response Authority (HERA) incubator in February 2021.[8] The emergency plan focused on the upscaling of existing vaccine production and adaptation of vaccines to virus variants, resulting in the formal launch of the HERA[9] in September 2021.

In addition, in recognition of the need for coordinated data sharing, the EC funded initiatives to fast track the development of the European COVID-19 Data Platform.[10]

Meanwhile, on the ground, the researchers working in the EC-funded COVID-19 projects related to data sharing observed that the unified and decisive response to the pandemic - in terms of the emergency funding mechanisms — could not always be matched when translated to the implementation phase. Delays in the authorisation of clinical trials and studies by national regulatory bodies of up to six months in several EU Member States, difficulties in engaging hospitals to participate in observational studies and a general lack of consensus in the interpretation of the General Data Protection Regulation (GDPR) were soon identified as barriers hampering project progress.[11,12] Data protection challenges related to compliance with GDPR and additional national and international regulations when sharing data and samples across borders, coincided also with numerous technical difficulties in sharing data across different institutional platforms, even within countries, due to the heterogeneity in the definitions of clinical events and epidemiological outcomes, and a lack of well-established data science protocols.[13]

However, these challenges were not new to international, multicentred clinical research projects conducted pre-COVID-19. For example, in projects like ZIKAlliance[14] and COMBACTE[15] the EU had already identified harmonisation and sharing of data in the health sector as an area requiring additional research and subsequently had invested in research projects such as ReCoDID[16] and SYNCHROS,[17] whose activities explicitly concentrated on these challenges. There were also a number of investments in pan-European infrastructures created for the purpose of accessing and sharing health data.[18,19] The EU Joint Action TEHDAS has been developing the principles and recommendation for the European Health Data Space (EHDS) which was formalised in the proposed EU Regulation of May 3rd 2022.[20] The EHDS is an important step in the right direction to overcome legal and ethical challenges for data sharing.[21] It explicitly includes cross-border access, emphasizes the importance of interoperability, and lays out a governance framework for data sharing of electronic health data. Though primarily focussed on electronic health records (EHR), it also includes clinical trials, research cohorts and biobank data within the 'secondary use of electronic health data' (chapter IV).

Despite these promising initiatives, the urgency to overcome the hurdles, delays and challenges in data sharing has never been so keenly felt than during the current pandemic, where the race for knowledge became a question of life or death.

Overall challenges in sharing health data

When appropriately used, evidence generated from diverse health data sources increases robustness and generalizability of research findings. However, as evidenced by the pandemic and before, effective health data sharing is hindered by important challenges including:

- *Regulatory frameworks for data protection across the EU, and elsewhere, do not facilitate data sharing.*

The GDPR sets the foundation for data protection across the EU, with specific requirements for health data, being considered sensitive and subject to stricter norms. Furthermore, EU members' own legislation can introduce additional layers of protection, increasing the complexity and the time needed for drawing up agreements between data sharing partners. If non-EU partners also intend to share or use data within a health data consortium, alignment needs to be sought between international data protection legislation and GDPR. This heterogeneous framework must be mapped out before any data sharing activity can start, and it may take a substantial amount of time, especially for large, multinational consortia, which are best placed to rapidly produce sound and robust evidence.

- *There is a lack of consensus on the choice of interoperability standards.*

There is a variety of community-developed standard terminology (e.g., CDISC/CDASH, OMOP, FIHR, SNOMED, LOINC etc.) for research studies as well as for health data in general, and similarly a wide range of IT solutions to handle these data. The heterogeneity of data sources (clinical trials, observational cohorts, patients' records, -omics, etc), and data collection tools calls for a meta-harmonization of common data and metadata standards and standardized processes to facilitate rapid data integration and use. Some initiatives are worth mentioning in this regard: 1) the Maelstrom Research cataloguing toolkit is a comprehensive and user-friendly web-based metadata catalogue based on existing catalogues and standards facilitating the interpretation and analyse of cohort data[22]; and 2) the CINECA project explores existing data representation as variables recorded, variable values and coding systems used in ten cohorts to construct a common minimal metadata model aligned with output from international standard activities[23] Although the FAIR data principles (Findable, Accessible, Interoperable, Reusable) are well known to data scientists, their adaptation in the medical world remains poor. Furthermore, there is a clear lack of robust, and trustful data sharing infrastructures that can streamline these standardized protocols, trace data management and thus ensure accountability.

- *Suboptimal data science literacy in the health sector.*

To achieve the above referenced standards, interdisciplinary collaboration in the medical field and digital literacy as well as technical skills of professionals working across the data sciences cycle, is needed. The complexity of the data and the intricate legislation covering data protection need highly qualified data scientists, IT engineers, and legal experts, as well as increased capacity among health care providers on data management to enhance the medical data quality, availability and accessibility.

Data sharing in observational studies

Approximately 95% of more than 230,000 primary studies published so far on SARS-CoV2 infection are indeed of an observational nature.[24] However, lower quality scores compared to non-SARS-CoV-2 papers were frequently found, even in the highest impact medical journals with a common issue being the very limited sample size.[25,26]

Data collection and harmonisation across different cohorts may be the solution to reach adequate statistical power to test several hypothesis and control for relevant biases, but it is not a straightforward task especially when across centres data collection and analysis is not planned in advance. Different sites may have important differences in data collection methodologies leading to sources of heterogeneity that might cause important biases in integrated results.[27] Proper harmonisation allows data to be comparable and, as a final goal, reliable and valid for integrated research analysis. Initiatives such as Maelstrom Research[28] have provided data harmonisation guidelines that ensure data quality, reproducibility and transparency of the process.[29] Unfortunately, the process of harmonising and integrating data from existing projects is often poorly documented.[30] Human resources and economic efforts should be made available to support the development and implementation of international standards and protocols for harmonisation.[31] Future strategies for large-scale harmonisation of individual data are needed to address the current limits caused by time-consuming manual work. Additionally, appropriate statistical methods should be employed to control for competing risks and other relevant biases that might arise even when harmonizing data from heterogeneous populations and including longer follow-up periods.

Cases in point — experiences from selected EU projects

Several multi-national EU-funded research projects address data harmonisation and data exchange, among them **ReCoDID** ('Reconciliation of Cohort Data for Infectious Diseases'),[16] **ORCHESTRA** (Connecting European cohorts to increase common and effective response to SARS-CoV-2 pandemic),[5] **unCoVer** (Unravelling data for rapid evidence-based response to COVID-19)[6] and **SYNCHROS** (SYNergies for Cohorts in Health: integrating the Role of all Stakeholders).[17] All four projects have been funded at different times by the European Union's Horizon 2020 research and innovation program (See Table 1 for an overview).

The **ReCoDID** project has been tasked with providing an harmonisation pipeline for COVID-19 cohorts in Europe, collaborating with the European COVID-19 Data Portal hosted by the European Molecular Biology Laboratory (EMBL).[32] ORCHESTRA exhibits important synergies with ReCoDID as both projects aim at the creation of pan-European cohort data sets for COVID-19, integrating data from existing and new large-scale cohort data sets across Europe. However, while the main objective of ReCoDID is to create a dedicated cohort data repository and significant efforts were invested in the legal and technical barriers to achieving this, ORCHESTRA focuses on the implementation of standardised new data collection among different cohort data to generate high quality evidence to improve the prevention and treatment of COVID-19. Indeed, more than 1,300,000 individuals from four broad cohorts (COVID-19 patients, general population, fragile population, and healthcare workers) have been enrolled in the ORCHESTRA cohort to date.

The two projects also differ with respect to their approach to centralised vs. federated data infrastructures, with ReCoDID adopting a centralised infrastructure and ORCHESTRA using primarily a federated infrastructure. **unCoVer**, a Coordinating and Support Action, also opted for a federated data infrastructure with a focus on data mostly from EHRs. These records cover over 22,000 hospitalised COVID-19 patients, as well as national surveillance and screening data and registries with approximately 2,000,000 COVID-19 patients. Accessing these records in the initial project phase was complicated by ethical considerations as is discussed later in this paper. With **SYNCHROS**, which was also a Coordinating and Support Action, the focus has been on the methodologies concerning cohort data comparability. The cohort mapping process that was conducted by SYNCHROS revealed a lack of standardized reporting when it comes to detailed information about cohort samples and data, and cohort harmonisation procedures. Indeed ORCHESTRA experienced first hand this hurdle: over 2,500 SARS-CoV-2-related variables were collected and had to be linked to unique international standard terminology codes provided by organizations such as SNOMED CT,[33] LOINC,[34] ATC,[35] and ICD-10.[36] But for some data elements a corresponding international code was not always found. Therefore, new concepts have been submitted to the most pertinent standard organisations to reinforce future global exchange of data and build a model to increase comparability of data in preparedness plans.[37]

Centralised vs. federated approaches

Centralised data repositories are effective solutions for the hosting and curation of large data sets, but important regulatory and ethical challenges have recently become evident. Ideally, standard language for broad informed consent that includes specifications about future use of data and transboundary sharing of data would enable centralised data repositories in Europe. This might be feasible for (observational) research studies, but for electronic health records from routine medical charts this is very challenging.

Recently, initiatives are emerging that attempt to address the inherent regulatory complexities and privacy concerns, using a federated data analysis and machine learning approach.[38] With these approaches, where individual-level data never leave the institution but are analysed locally in a parallel way, only aggregated information is communicated to the central node and later iteratively integrated.

Indeed, since ORCHESTRA includes cohorts already established prior to the start of the project, for many cohorts the participant-level data cannot be directly shared. To enable the application of the agile and advanced analysis methods avoiding direct data sharing, an infrastructure for federated data analysis and machine learning network has been established based on OPAL-DataSHIELD.[39] This infrastructure enables conducting participant-level analysis without revealing/exchanging the participant-level data, as only the computed parameters describing the entire datasets are communicated to the analyst. The network consists of the access point hosted by the main institution and of the local nodes maintained by the data-owners, such as hospitals. Such design also supports parallelization — since the computations can run simultaneously across local nodes. Although at the moment the Opal-DataSHIELD[39] does not support a complete landscape of the analysis methods and machine learning tools, it seems the best option for the datasets that cannot be directly shared. The largest bottleneck of this process is (i) the informatics resources, namely the server infrastructure,

Project Title	Acronym	Project Objectives	Start date	Duration	EU Funding	Approx. % dedicated to data harmonisation	N° of scientific publications[a]
Reconciliation of Cohort Data for Infectious DIseases	ReCoDID	- To develop an integrated sustainable platform for collating data within and across infectious disease cohorts that facilitates the use of CE and HDL data for detection, treatment, and prevention of infection of known and unknown pathogens - To reconcile frequently encountered barriers to sharing data and human specimens, and create innovative solutions for shared ownership, linked data and biorepositories, and collaborative analysis	Jan 1st 2019	4 years	€7.760.021	60%	35
Connecting European cohorts to increase common and effective response to SARS-CoV-2 pandemic	ORCHESTRA	- To create a new pan-European cohort built on existing and new large-scale population cohorts in European and non-European countries of SARS-CoV-2 infected and non-infected individuals of all ages and conditions to assess risk factors, drivers of disease, and long term consequences -To develop evidence-based recommendations for effective prevention, protection and optimized treatment of COVID-19 patients with a special focus on 'at risk' population - To assess impact of environmental factors, socio-economic determinants, lifestyle and confinement measures on the spread of COVID-19 - To provide a model for data collection for responsiveness for future pandemic outbreaks	Dec 1st 2020	3 years	€27.887.638	10%	40
Unravelling Data for Rapid Evidence-Based Response to COVID-19	unCoVer	- To monitor, identify, and facilitate the access and use of COVID-19-related Real World Data (RWD) - To identify data gaps, and marginalized populations and proactively seek synergies with complementary existing and planned clinical databases related to COVID-19 - To provide a platform for the use of dissimilar data sources capable of streamlining ethical and legal aspects, by innovative computational resources - To bring together expertise on the use of advanced computational, epidemiological and biostatistical methods to handle heterogeneous, and multi-layered information	Nov 15th 2020	2 years	€2.997.440	70%	5
SYNergies for Cohorts in Health: integrating the ROle of all Stakeholders	SYNCHROS	- To establish a sustainable European strategy for the development of the next generation of integrated population, patient and clinical trial cohorts - To map the cohort landscape in Europe and large international initiatives (SYNCHROS Repository) - To identify best methods for integrating cohort data in order to enable the harmonisation of past and future data collection - To identify solutions for addressing practical, ethical and legal challenges in integrating data across patient, clinical trial and population cohorts.	Jan 1st 2019	3,5 years	€ 1.991.812	N/A Data harmonisation not a direct project objective	10

***Table 1*: Overview of projects.**

[a] Published and submitted for publication.

and (ii) the competency and understanding on the side of the data owners and local regulatory bodies such as ethics boards.

Similarly, in unCoVer, heterogeneous data are described, harmonised and integrated into a multi-user data repository operated through Opal-DataSHIELD,[39] an interoperable open-source server application. This federated infrastructure offers the most efficient and secure approach to handling highly sensitive patient information derived from EHR, information which has not been collected for research purposes and demanding a particularly secured environment as well as close monitoring of data protection compliance.[40] It should be noted that unCoVer also faced significant barriers from the ethics perspective in the initial project start-up phase when setting up the federated infrastructure and therefore delays in the planned analyses that could have helped alleviated pandemic outcomes.[40]

For ReCoDID, on the other hand, a centralised data curation pipeline hosted at the European COVID-19 Data Portal (via EMBL) is the preferred way forward due to the opportunity to merge large OMICS data with the clinical-epidemiological data and enable future large-scale investigations across different data sets and across the divide between high dimensional and well-characterized clinical-epidemiological data. However, federated data analysis alternatives are investigated, involving national hubs of the European Genome Archive infrastructure (EGA) by EMBL.

Re-thinking the research funding format

To date, EU research funding mechanisms have adhered to the classical format in terms of call generating procedure, call publication and grant competition by different competitive consortia. However, the pandemic has shown the limitations of this classical format, which can be slow to generate evidence and knowledge. In times of dynamic change, there is a need for generating evidence on the fly to enable policy makers to make scientifically informed decisions. Furthermore, by focusing solely on the competitive element between different consortia, some EU-regions have been continuously overlooked by past and running calls.

Therefore, it may be worthwhile to consider alternative formats of EC calls to foster the collaborative and networking aspects, also with respect to a more efficient approach of accessing and sharing data. This means that: (a) calls could include incentives for transdisciplinary consortia, comprising different fields of public health relevant topics, but also containing an incentive and formula for the inclusion of data owners; (b) geographic origin could be interpreted more broadly by encouraging and facilitating participation of different European regions instead of a single partner institution from Member States, thus fostering from the outset a wider landscaping of data sources; (c) standard wording for broad informed consent could be incorporated in call documents as a requirement to be adopted by future consortia in order to facilitate data sharing (d) dedicated funding could be set aside (or related separate calls could be initiated) for coordination and harmonisation, of multicentic studies and last but not least (e) knowledge scouting activities could be included as a dedicated work package producing evidence while studies are underway in order to confirm scientific importance and usefulness of early project results for decision making.

In fact, when the NIH published a call on Long COVID research in the spring of 2021, it actually published four related calls: (i) the central call aimed at prospective data collection in a multicentric cohort study, plus three related calls in supporting roles; (ii) a call for a 'clinical science core', (iii) a call for a 'data resource core' for data management, harmonization, and sharing, and (iv) a call for a 'biorepository core'. Separating out a 'clinical science core' from the actual cohort implementation could have the advantage of advancing more rapidly on the legal and ethical challenges.[41]

Roadmap for the future

The scientific community and stakeholders responded to the COVID-19 pandemic in an unprecedented manner in terms of collaborations, funding and data sharing. The existing molecular databases quickly adapted to facilitate the collection of data on the new virus, and various new platforms emerged for data curation and data sharing. COVID-19 patient cohorts were created with equal speed, but they lacked coordination and harmonisation, consequently crippling a deep and sensitive analysis.

High quality and detailed participant-level machine-readable clinical data are sensitive, and their usage should remain controlled to secure protection of privacy as well as ethical use. At the same time, a lack of well-defined guidelines for data sharing, knowledge of legal regulations or technical resources to standardise and share data often hinder scientific development. In addition, multiple community developed data standards are now available, making it difficult to select between them. Interoperability on a meta-level between these standards has not yet been achieved.

Large collaborative projects, such as RECODID, ORCHESTRA and unCoVer create a collaborative environment that enables scientific discovery and methods development. However, this does not entirely remove the need for local support regarding the legal and informatics aspects. Thus, strengthening relationships between the IT and computational experts in the local institutions, like hospitals or universities, as well as the resources for the IT personnel and infrastructure, could enable innovation in developing other federated approaches for the clinical-data analysis on a larger scale. This may also necessitate making funding

Critical Area	Suggested Actions
Inconsistency in application of GDPR across Member States	Non-binding implementing rules / Code of conduct recognised by Member States
Stringent local legal and ethical requirements impeding rapid collection of data and analysis	Guidelines (or Stewardship) for application of GDPR in case of pandemic or major public health threat
Lack of common standards on data use, and data interoperability	Mandate internationally endorsed standard terminologies and classifications in funding frameworks
Lack of agreement on the use of metadata standards	Catalogue the international use of metadata standards to empirically determine most common used standards; incentivize researchers for proper metadata documentation
Lack of standardised reporting on harmonisation procedures	Training and education on best practices in reporting harmonisation procedures/ High-quality peer-review on publication reporting harmonisation outcomes
Multiple community-developed standards for interoperability	Development of meta-harmonization tools for interoperability between community developed standards
Poor digital literacy and data science skills of staff of data owners (hospitals etc.)	Institutional capacity building for staff and resources for IT infrastructure and strengthening of inter-institutional collaboration.
Standard funding frameworks do not always adapt well to projects formulated to address a pandemic	Devise alternative formats with a focus on collaborative and network aspects favouring complementarity as much as competitiveness
Barriers of sharing individual patient data for EHR and for some retrospective cohort data	Further development and investment in federated learning and analysis networks and technology
Broad informed consent for data sharing often not available	Make broad informed consent for future use of data mandatory as part of funding frameworks (especially for observational cohort data) / Introduce standardised language for informed consent documents
Manipulation of data for pseudonymization/anonymization purposes may undermine its scientific value	Align any manipulation of data for pseudonymization purposes to the longitudinal characteristics of cohort studies concerned (i.e. temporal and location dimensions).
Retrospective harmonization is extremely labour-intensive	Dedicated funding for retrospective harmonization of valuable cohort data (selective)/ Investment in future development of (AI-based) harmonization routines which are less labour-intensive.

***Table 2*: Summary of limitations of current data sharing process and suggested actions to address these limitations.**

available for IT and data curation support for the data contributors who often are not full partners of the research consortia.

Furthermore, further efforts are needed from the individual Member States towards the adoption of common standards in terms of health data. Additionally, as foreseen also by the GDPR, unprecedented public health threats such as the COVID-19 pandemic, should prompt the re-evaluation of some very stringent local legal and ethical requirements that do not allow researchers to deliver informative research in a timely manner. Alternatively, Member States could provide common guidance on the application of such requirements in a pandemic situation.

Federated data analysis and machine learning does offer a solution for the delivery of a highly agile, robust and scalable system enabling fast real-time analysis of the highly sensitive up-to date clinical data. It could also enable integration of the clinical participant-level data with the complex often non-tabular-OMICS data. However, this approach reinforces the need for the use of the standardised data collection methods, as the analyst is not able to directly interact and view the data. Thus, as mentioned above, this solution requires the data owners to be equipped with the infrastructure and expertise to maintain the local nodes of the harmonised datasets. This requirement, in itself, poses a challenge because not all federated approaches are created equal since some systems rely on more resources than others. Furthermore, the fact that analyst cannot directly see the data, prompts the question about their ability to identify the source of potential distortions.

Centralised data infrastructures are still important and — where data can be shared centrally (e.g., when broad informed consent for future use is available) — may be the preferred option because of the ease of curation / storage and the need for computational resources for OMICS data analysis. While the federated approach is further developed, the centralised approach will remain an important alternative, especially if the data is shared between trusted institutions with the appropriate safeguards in place.

An important final consideration that concerns both centralized and federated arrangements relates to the limits of anonymization conversions: data cannot be converted in any formats and anonymized indefinitely and therefore pseudonymization is used. The problem, as emerged from the SYNCHROS stakeholders dialogue, is that the legal basis for such pseudonymization conversions is still unclear. Moreover, any manipulation of data for pseudonymization/anonymization purposes may undermine its scientific value. An important way forward is thus to ensure any manipulation of data for pseudonymization purposes will be aligned to the longitudinal characteristics of cohort studies concerned (i.e. temporal and location dimensions).

The recent proposal for a Regulation the European Parliament and of the Council for a European Health Data Space (EHDS)[20] offers a promising framework to address many of the challenges addressed in this paper, where the future EHDS regulation would provide the much needed legal basis to access and use secondary health data under GDPR (for research, innovation, policy making, patient safety or regulatory activities) and where the EHDS would be built on strong data governance, data quality and interoperability.

It will be important to monitor the evolution of this important piece of legislation and its practical application by EU Member States according to the provisions foreseen (e.g., the setting up of a digital health authority at MS level; a common infrastructure at EU level to facilitate cross border exchange of electronic health data; self-certification schemes for EHR systems to ensure interoperability; the secure processing environment for secondary data and related costs; and provisions related to the joint controllership for EU infrastructures.)

A summary of the main critical issues faced in the implementation of the projects and the proposed solutions to address these issues are listed in Table 2. With the limit of relying on the authors' personal opinion and experience, and in view of the evolving legislative framework, we believe that this combined project overview offers numerous areas for reflection and confirms that continued efforts are required not only on the part of Member States but also at the EU level to provide certainty and clarity when it comes to the practical, methodological, ethical and legal aspects of data sharing and data harmonisation.

Funding

The ReCoDID project has received funding from the European Union Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825746, and is supported by the Canadian Institutes of Health Research, Institute of Genetics (CIHR-IG) under Grant Agreement No. 01886-000.

The ORCHESTRA project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101016167.

The unCoVer project is funded by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 101016216.

The SYNCHROS project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 825884.

Contributors

ET and AWF conceived the idea of the manuscript as part of the ORCHESTRA project. All authors contributed to the first draft of the manuscript. RJD, AG and EC reviewed and edited the final paper. All authors read and approved the final paper.

Declaration of interests

MB reports grants to UMCU from Janssen Vaccines, Novartis and CureVac, consulting fees from Astra Zeneca, Pfizer, Janssen Vaccines, Novartis,Takeda,Janssen Vaccines and payments from Sanofi for Participation on a Data Safety Monitoring Board or Advisory Board. ET, EC and RJD report H2020 funding for ORCHESTRA. JH reports H2020 funding for ORCHESTRA, funding from the German Research Council for the SEPAN Project, funding from German Ministry of Education and Research for the EMUNE and the INSIDe Project and funding from Volkswagen Stiftung for the E2 project. AS and EV report H2020 funding for SYNCHROS. TJ reports H2020 funding for ReCoDID. All other authors declare no competing interests.

References

1 Advancing knowledge for the clinical and public health response to the 2019-nCoV epidemic. EU Commission; 2020. https://cordis.europa.eu/programme/id/H2020_IBA-SC1-CORONAVIRUS-2020-3. Accessed 16 March 2022.
2 FIRST "ERAvsCORONA" ACTION PLAN, working document, 7 April 2020. https://ec.europa.eu/info/sites/default/files/covid-first eravscorona_actions.pdf. Accessed 22 March 2022.
3 Innovative and rapid health-related approaches to respond to COVID-19 and to deliver quick results for society for a higher level of preparedness of health systems. EU Commission; 2020. https://ec.europa.eu/info/events/2nd-special-call-expression-interest-respond-coronavirus-information-session-2020-may-20_en. Accessed 16 March 2022.
4 RECOVER. Rapid European COVID-19 Emergency Response research. https://www.recover-europe.eu/. Accessed 16 March 2022.
5 ORCHESTRA. Connecting European international cohorts to increase common and effective response to SARS-CoV2 Pandemic. https://orchestra-cohort.eu/. Accessed 10 March 2022.
6 unCoVer. Unravelling Data for Rapid Evidence-Based Response to COVID-19. https://uncover-eu.net/. Accessed 11 April 2022.
7 EU-RESPONSE. European research and preparedeness network for pandemic and emerging infectious diseases. https://eu-response.eu/. Accessed 18 April 2022.
8 HERA Incubator, Anticipating together the threat of COVID-19 variants, Brussels 17.02.2021) (European bio-defence preparedness plan against COVID-19 variants), *European Commission*, 78, final; https://ec.europa.eu/info/sites/default/files/communication-hera-incubator-anticipating-threat-covid-19-variants_en.pdf.
9 Introducing HERA, the European Health, Emergency preparedness and Response Authority, the next step towards completing the European Health Union, Brussels, 16.9.2021, *European Commission*, 576, final; https://health.ec.europa.eu/system/files/2021-09/hera_2021_comm_en_0.pdf.
10 COVID-19 Data Portal. https://www.covid19dataportal.org/. Accessed 24 March 2022.
11 Diallo A, Trøseid M, Simensen VC, et al. Accelerating clinical trial implementation in the context of the COVID-19 pandemic: challenges, lessons learned and recommendations from DisCoVeRy

and the EU-SolidAct EU response group. *Clin Microbiol Infect.* 2022;28(1):1–5.
12 Goossens H, Derde L, Horby P, Bonten M. The European clinical research response to optimise treatment of patients with COVID-19: lessons learned, future perspective, and recommendations. *Lancet Infect Dis.* 2022;22(5):e153–e158.
13 European Commission, Directorate-General for Research and Innovation, Maxwell L. Maximising investments in health research: FAIR data for a coordinated COVID-19 response: workshop report, 2022. https://data.europa.eu/doi/10.2777/726950.
14 ZIKAlliance. A global alliance for zika Virus control and prevention. https://zikalliance.tghn.org/. Accessed 16 March 2022.
15 COMBACTE. Combatting bacterial resistance in Europe. https://www.combacte.com/. Accessed 16 March 2022.
16 ReCoDID. Reconciliation of Cohort data in Infectious Diseases. https://recodid.eu/. Accessed 17 April 2022.
17 SYNCHROS. SYNergies for Cohorts in Health: integrating the ROle of all Stakeholders. https://synchros.eu/. Accessed 1 April 2022.
18 https://www.ema.europa.eu/en/about-us/how-we-work/big-data/data-analysis-real-world-interrogation-network-darwin-eu. Accessed 17 July 2022.
19 The European Health Data & Evidence Network. https://www.ehden.eu. Accessed 16 March 2022.
20 Proposal for a regulation of the European Parliament and of the Council on the European Health Data Space. COM(2022) 197 final. Strasbourg, 3.5.2022. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0197&from=EN. Accessed 17 July 2022.
21 Genovese S, Bengoa R, Bowis J, et al. The European health data space: a step towards digital and integrated care systems. *J Integrated Care.* 2022. https://www.emerald.com/insight/content/doi/10.1108/JICA-11-2021-0059/full/html.
22 Bergeron J, Doiron D, Marcon Y, Ferretti V, Fortier I. Fostering population-based cohort data discovery: the Maelstrom research cataloguing toolkit. *PLoS One.* 2018;13(7):e0200926.
23 Thorogood A, Rehm HL, Goodhand P, et al. International federation of genomic medicine databases using GA4GH standards. *Cell Genom.* 2021;1(2):100032.
24 Epistemonikos. https://www.epistemonikos.org/. Accessed 17 July 2022.
25 Jung RG, Di Santo P, Clifford C, et al. Methodological quality of COVID-19 clinical research. *Nat Commun.* 2021;12(1):943.
26 Zdravkovic M, Berger-Estilita J, Zdravkovic B, Berger D. Scientific quality of COVID-19 and SARS CoV-2 publications in the highest impact medical journals during the early phase of the pandemic: a case control study. *PLoS One.* 2020;15(11):e0241826.
27 Granda P, Blasczyk E. *Data Harmonization Guidelines for Best Practice in Cross-Cultural Surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan; 2010.
28 Maelstrom. https://www.maelstrom-research.org/. Accessed 17 July 2022.
29 Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol.* 2017;46(1):103–105.
30 Rodríguez-Laso Á, Rico-Uribe LA, Kubiak C, Haro JM, Rodríguez-Mañas L, Ayuso JL. A map of the initiatives that harmonize patient cohorts across the world. *Front Public Health.* 2021;9:1–12.
31 Hofer SM, Piccinin AM. Integrative data analysis through coordination of measurement and analysis protocol across independent longitudinal studies. *Psychol Methods.* 2009;14(2):150.
32 European Molecular Biology Laboratory EMBL. https://www.embl.org/. Accessed 1 July 2022.
33 El-Sappagh S, Franda F, Ali F, Kwak K-S. SNOMED CT standard ontology based on the ontology for general medical science. *BMC Med Informat Decis Making.* 2018;18(1):1–19.
34 McDonald CJ, Huff SM, Suico JG, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin Chem.* 2003;49(4):624–633.
35 WHO Collaborating Centre for Drug Statistics Methodology, *Guidelines for ATC classification and DDD assignment 2021*, World Health Organization; Oslo, Norway, 20.
36 Fung KW, Xu Julia, Bodenreider O. The new international classification of diseases 11th edition: a comparative analysis with ICD-10 and ICD-10-CM. *J Am Med Informat Assoc.* 2020;27(5):738–746.
37 Rinaldi E, Stellmach C, Rajkumar NMR, et al. Harmonization and standardization of data for a pan-European cohort on SARS- CoV-2 pandemic. *npj Digital Med.* 2022;5(1):75.
38 Marcon Y, Bishop T, Avraam D, et al. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *PLoS Computat Biol.* 2021;17(3):e1008880.
39 DataSHIELD. Secure Bioscience Collaboration. https://www.datashield.org. Accessed 1 July 2022.
40 Peñalvo JL, Mertens E, Ademović E, et al. Unravelling data for rapid evidence-based response to COVID-19: a summary of the unCoVer protocol. *BMJ Open.* 2021;11(11):e055630.
41 NIH. Research Opportunity Announcement OTA-21-015A Post-Acute Sequelae of SARS-CoV-2 Infection Initiative: Clinical Science Core, Data Resource Core, and PASC Biorepository Core 2022. https://covid19.nih.gov/sites/default/files/2021-02/PASC-ROA-OTA-3-Cores.pdf. Accessed 16 March 2022.