Title

FAIR, ethical, and coordinated data sharing for COVID-19 response: a review of COVID-19 data sharing platforms and registries

Authors

Lauren Maxwell[1], Priya Shreedhar[1], Delphine Dauga[2], Peter McQuilton[3], Robert F Terry[4], Alisa Denisiuk[5], Fruzsina Molnar-Gabor[6], Abha Saxena[7], Susanna-Assunta Sansone[3]

Universitätsklinikum Heidelberg, Heidelberger Institut für Global Health, Heidelberg, Germany.[1] Bioself Communication, France.[2] Oxford e-Research Centre, Department of Engineering Science, University of Oxford, UK.[3] The Special Programme for Research and Training in Tropical Diseases, World Health Organization, Geneva, Switzerland.[4] Georg-August-University, Göttingen, Germany.[5] Heidelberg Academy of Sciences and Humanities, Heidelberg, Germany.[6] Independent Bioethics Advisor, Geneva, Switzerland.[7]

Abstract

Data sharing is central to the rapid translation of research into advances in clinical medicine and public health practice. In the context of COVID-19, there has been a rush to share data, marked by an explosion of population- and discipline-specific resources for collecting, curating, and disseminating participant-level data. We present a comprehensive overview of COVID-19-related platforms and registries that harmonize and share participant-level clinical, OMICs, and imaging data and metadata, and describe how these initiatives map to best practice for ethical, equitable, and FAIR management of data resources. Data sharing resources were concentrated in high income countries and siloed by comorbidity, body system, and data type. Resources for sharing clinical data were less FAIR than those for sharing OMICs or imaging data. We review gaps and redundancies in COVID-19 data sharing efforts and outline recommendations to build on existing synergies and align with frameworks for effective and equitable data reuse.

Summary for dissemination activities

Understanding how data are shared can help funders and researchers identify gaps and redundancies to improve global collaboration in the research response to COVID-19. In this manuscript, we present an in-depth overview of the ever-expanding universe of COVID-19-related platforms and registries for sharing clinical, OMICs, and imaging data and review how these initiatives map to best practice for ethical, equitable, and effective data sharing and application of the FAIR Principles for managing data resources. In large part, registries and platforms for improving the availability harmonized clinical or epidemiological, participant-level COVID-19 data had not adopted community developed standards for participant-level data and were most often siloed by

data type, comorbidity, body system, and population type. To better respond to the ongoing pandemic, we need to move from fragmented, overlapping and competing data sharing efforts to a coordinated nexus of interconnected, longitudinal, participant-level data.

Introduction

There is myriad public health, ethical, economic, and scientific arguments for collecting, harmonizing, and sharing public health-related, participant-level data from research studies, disease surveillance systems, and routine clinical care. These include fast tracking the development and evaluation of preventative measures, diagnostics, and treatments; avoiding the human and economic cost of unnecessary research; and more effectively distinguishing between clinically relevant and spurious sources of heterogeneity to optimize prevention and treatment measures for diverse populations. The urgency of the ongoing COVID-19 pandemic has foregrounded the importance of data sharing, in some cases, pitting collaborators or similar initiatives against one another in the quest for funding and data producers to support data sharing activities.

Many researchers share data through uploading their datasets to data lakes or dataverses, data storage and sharing resources, like GitHub, where the data are not harmonized at the participant level and there are few-to-no restrictions on the types of studies from which data can be shared. In this review, we focused on identifying and describing COVID-19-related platforms and registries, data sharing resources that conduct pro or retrospective harmonization of participant-level data. In most cases, registries require data contributors to upload data using a shared case report form (CRF) while researchers can upload datasets with different data dictionaries to data sharing platforms. Both data sharing platforms and registries may limit eligibility to certain types of data or populations. Data sharing platforms generally represent greater investments because of the diverse inputs needed for retrospective harmonization, their focus on high- rather than low-dimensional data types (e.g. OMICs, including human and pathogen genomic data, human metabolomic data, etc., and imaging rather than clinical data), and more expansive inclusion criteria, which allow for the collection of a greater volume and diversity of data (see Supplementary Table 1 for working definitions of data sharing resources).

Collecting participant-level data and descriptive metadata and harmonizing and sharing participant-level data are resource intensive activities that require expertise in physiology, diagnostics, the trajectory and etiology of infection, risk factors and comorbidities, standards for the interoperability of meta- and participant-level data, harmonization, data sharing-related laws, research ethics, and community engagement. In addition to concerns about maximizing data sharing investments through fostering the interoperability of related platforms and registries, the rush to facilitate COVID-19-related data sharing through the extension of existing platforms and the establishment of novel registries raises a number of questions related to how data sharing efforts map to the FAIR principles for data resources[1] and best practice for the ethical reuse of participant-level data[2,3].

To explore these and other questions, we collected data on a number of domains of interest for evaluating how resources for collecting, harmonizing, and sharing participant-level COVID-19 data and related metadata correspond to frameworks for public health-related data sharing, including the Global Health Security Initiative and Global Research Collaboration for Infectious Disease Preparedness (GloPID-R) Principles of Sharing Data in Public Health Emergencies[4], COVID-19 National Core Studies (NCS)Data Sharing Principles[5], Global Alliance for Genomics and Health (GA4GH) Framework for responsible sharing of genomic and health-related data[6], and the CARE Principles for Indigenous Data Governance[7].

Methods

We conducted a monthly search of Google and Google Scholar between May 2020 and June 2021 using text terms for COVID-19 and for data sharing resources (Supplementary Note 1) to identify relevant platforms and registries that collect, harmonize, and share COVID-19-related participant-level clinical, human or pathogen OMICs, and high dimensional imaging data. To account for English-language bias in the search strategy, we contacted investigators that work on COVID-19-related data sharing in Asia, Africa, and Latin America and applied natural language processing (NLP) to the Covid-19 Open Research (CORD-19) Dataset[8] in March 2021 to identify additional data sharing resources (Supplementary Note 2).

We consulted with end users of harmonized, participant-level COVID-19 data from different fields to identify information that would be useful for them to evaluate the utility of different data sharing resources. We collected general information on the resource (e.g. lead organization, location, funding), linkages between data types at the participant-level, resource metrics for success (e.g. number of dataset uploads and downloads), criteria used to evaluate resource adherence to the FAIR principles and the outcomes of those evaluations, data access mechanisms and governance structure, deidentification of data, ethics review and broad consent related requirements, community engagement, and benefit sharing with data contributors and source communities. We developed a REDCap[9] questionnaire (Supplementary Note 3) to collect required information and distributed the survey to 31 data platform or registry teams where required information was not provided on the resource website. Following four months of bimonthly reminders, 18 of the 31 data sharing resources we contacted completed the online survey. Where survey responses contrasted with information available online, we used the survey data.

How FAIRness is evaluated depends on the data type and community-specific needs and preferences. When the FAIR principles were first published[1], they were necessarily aspirant and vague. Over time, different interpretations and extensions of the principles have developed, alongside a number of assessment tools (listed at https://fairassist.org). We conducted a qualitative evaluation of registry and platform adherence to the FAIR principles using four basic criteria: (1) whether the resource was discoverable via a persistent identifier (PID); (2) whether information on how to access data was available on the resource website; (3) whether the resource implemented a community-developed standard for participant-level data or metadata; and (4) whether the resource specified a data usage license or agreement. We conducted a quantitative evaluation of how registries for sharing participant-level, clinical data align with the FAIR principles through applying the FAIRshake algorithm[10] to a set of criteria that we identified as most important for evaluating the utility of these resources (Supplementary Note 4). We adapted existing criteria for our specific use case through a combination of a manual review of the FAIR maturity indicators[11] and the Research Data Alliance (RDA) FAIR Data Maturity model output[12] and a review of the algorithms used by semi-automated tools, including FAIRshake[10], FAIR evaluator[11], and FAIR-checker[13]. All figures were created in Tableau Desktop 2021.2 with the exception of Figure 2, produced through open source code on FAIRsharing.org.

Data availability

The dataset describing the 68 platforms and registries that collect, harmonize, and share COVID-19 related participant-level clinical, OMICs, and/or imaging data and an additional 13 meta repositories that share or otherwise facilitate access to COVID-19-related datasets or data sharing resources is available for comment on Zenodo (https://zenodo.org/record/5101817#.YRMHx44zaUk). Twenty-eight of these resources, which provide information on registry contacts, license, support information or data accessibility conditions, and where the data, although not public, are accessible on request and/or published in a scientific paper, and/or shared as report or dashboard were further described in FAIRsharing[14], a global resource that interlinks databases, standards and policies. Metadata for these resources are available on FAIRsharing via a dedicated collection: https://fairsharing.org/collection/TDRCOVID19Participantleveldatasharingplatformsregistries

Results

We identified 47 registries and 21 platforms that collected, harmonized, and, in some cases, shared participant-level COVID-19 human subjects' data. All but two of these were identified through the monthly Google searches rather than the NLP approach (see Supplementary Table 2 for citations identified through the NLP strategy). COVID-19 data sharing resources were overwhelmingly data type specific. Almost all registries (45 of 47) were limited to clinical data; two included clinical and high-dimensional imaging data. Eleven of the 21 platforms included OMICs data and six included high dimensional imaging data (e.g. CT scans). Nine platforms included more than one data type.

Long COVID affects approximately 30–87% of adults[15,16] and 5–8% of children[17,18] who are infected with SARS-CoV-2 and harmonized, longitudinal datasets with linked clinical, human and pathogen OMICs, and imaging data may facilitate long COVID-related prognosis and treatment and help identify participant-level factors correlated with the emergence of variants of concern (VOC) and VOC-related differences in etiology and vaccine efficacy. Supplementary Figure 1 shows registry- and platform-specific participant-level linkages between data types. About a third of platforms (N=9) and half of the registries (N=25) included longitudinal clinical data. While no registries included human or pathogen OMICs data, six platforms included longitudinal human OMICs data; two of those six also included longitudinal imaging data. Four platforms, (CanCOGeN, N3C, dbGAP, ReCoDID), included linked longitudinal clinical and human and pathogen OMICs data. Only one platform (N3C) included longitudinal data on all datatypes, including clinical, host and pathogen OMICs, and high dimensional imaging data.

Most registries (N=44), but only a few platforms (N=4) limited data to populations with a particular coinfection, comorbidity, assessment, treatment, or outcome of interest. There were several instances of registries that covered the same comorbidities, including six registries for different forms of cancer, four registries for blood conditions, four registries related to cardiovascular system diseases, seven registries for skin conditions, three registries for rheumatic disease, three registries for issues related to the digestive system, two registries for liver disease, two registries for neurological conditions, and two registries for diabetes. An additional three registries were limited to individuals with kidney disease, multiple sclerosis, and patients receiving extracorporeal membrane oxygenation (ECMO). Several registries collected data on pediatric (N=4) or pregnant (N=2) populations. Of those, two registries included data on pediatric cancer patients and one on pediatric patients with rheumatic disease. All platforms and two-thirds of registries (N=31) included data from participants of all ages. Nine registries were limited to data on adults aged 18 and over.

Supplementary Figures 2 & 3 show the global distribution of platforms and registries for sharing participant-level, COVID-19-related data. One third of platforms (N=7) and 59% of registries (N=28) were based in the US; 57% of platforms (N=12) and 36% of registries (N=17) were based in Europe; and one registry was based in Brazil and in Israel, respectively. Most platforms and registries (N=17, 71%, N=28, 59%, respectively) accepted data from any country; six platforms (28%) and 19 registries (40%) were country or region specific.

For resources that collected clinical data, most registries (N=43; 91%), but only one platform, were limited to prospective harmonization of participant-level data through a shared electronic case report form (eCRF). Four platforms (19%) and two registries (4%) conducted both prospective and retrospective harmonization; one registry for clinical data conducted only retrospective harmonization (ACR CIRR). Most registries and platforms that included prospective harmonization of clinical data provided a REDCap-based eCRF (29 of 54 platforms and registries; 54%). Other registries and platforms used Qualtrics (N=2), SurveyMonkey (N=2), OpenApp (N=2), or QMENTA (N=1) data capture software. At the beginning of the epidemic, the World Health Organization (WHO), International Severe Acute Respiratory and Emerging Infection Consortium (ISARIC), and Infectious Diseases Data Observatory (IDDO) created an open access series of REDCap-based eCRFs which applied CDISC's SDTM standards (https://doi.org/10.25504/FAIRsharing.s51qk5). Of the 54 resources that included clinical data, only one platform (IDDO) and registry (CAPACITY) reported using the WHO/ISARIC/IDDO eCRFs, which represents a missed opportunity for prospective harmonization in COVID-19 response.

We present an overview of how COVID-19-related resources for collecting, harmonizing, and sharing participant-level data map to the FAIR principles and best practice for ethical and equitable data sharing in Table 2 (see Supplementary Table 3 for related text from each set of principles). The FAIR principles focus on the machine-actionability of data and related metadata, findability, accessibility, interoperability, and reusability[1]. The quantification of how registries for sharing participant-level clinical data map to the FAIR criteria is presented in Supplementary Table 3 with the important caveat that quantitative evaluations are at an exploratory stage. The community continues to work on harmonizing the algorithms used to evaluate the application of FAIR indicators across disciplines as many tools for quantifying FAIRness yield divergent results. Therefore, we focus our discussion on the results of the qualitative evaluation of FAIRness, using the four main criteria described earlier. We considered resources that met none or one of the criteria as not very FAIR and resources that met two or more criteria as FAIR enough. As shown in Figure 1a, platforms were generally more closely aligned with the FAIR principles than registries and resources that were comorbidity or population specific. As indicated in Figure 1B, registries and platforms for harmonization of clinical or epidemiological data were much less FAIR than those that included high dimensional data types.

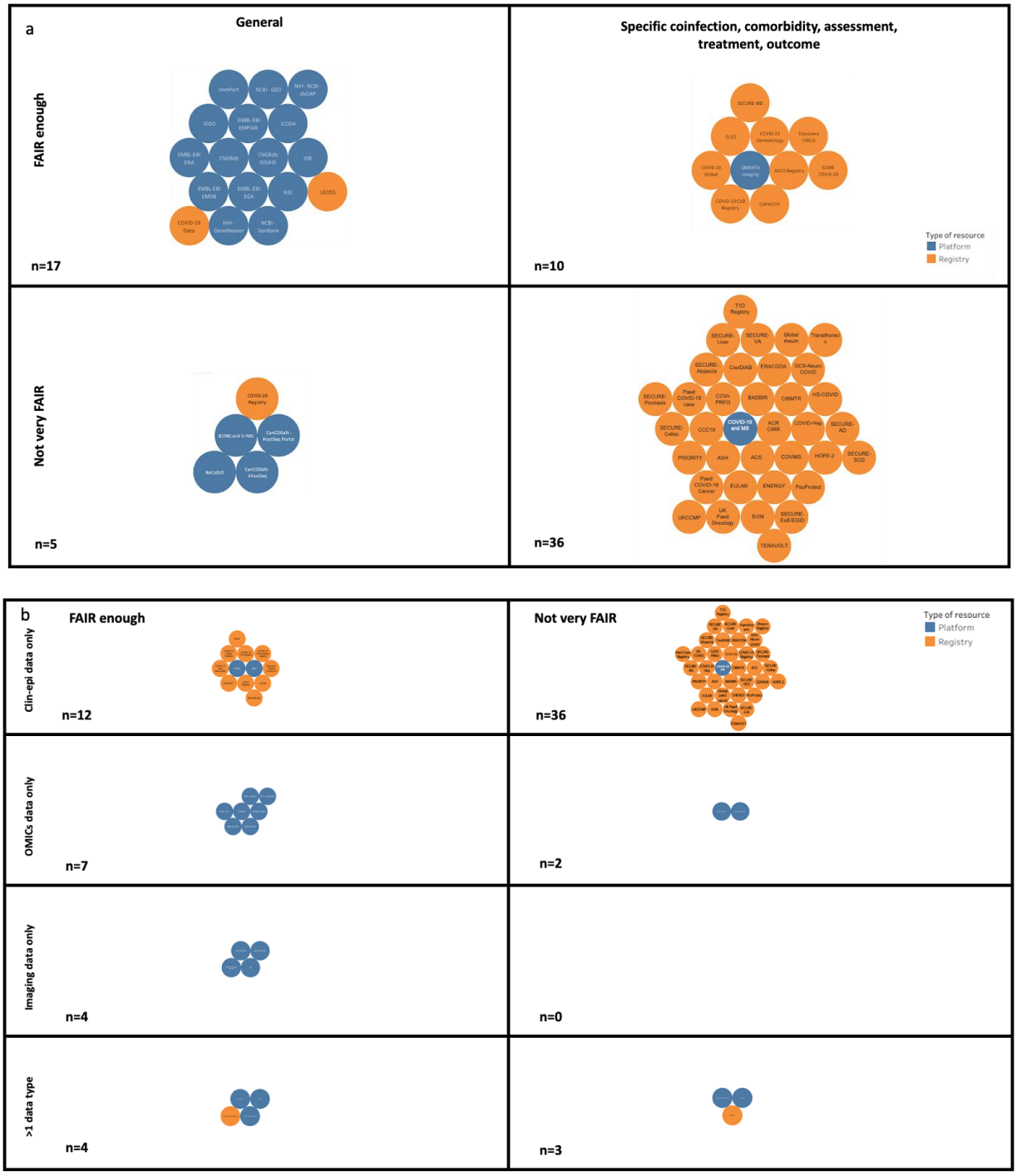


Fig.1 Qualitative evaluation of FAIRness. **a**. Disease-specific platform and registry correspondence with the FAIR criteria for data resources. **b**. Participant-level data types hosted by platforms and registries and correspondence with the FAIR criteria for data resources.

Our evaluation, which included the registration and curation of eligible resources identified through our search in FAIRsharing.org, improved the FAIRness of a number of platforms and registries, through improving discoverability and availability of descriptive metadata. Specifically, a digital object identifier (DOI) was assigned to two platforms and 18 registries, which did not have a PID;

which is central to findability. Additionally, we recorded information on the data accessibility mechanism and terms of use, elements essential to accessibility on Fairsharing.org. Lastly, we collected and recorded information about the data and metadata standards used by these resources; standards are fundamental to interoperability and reusability.

Community-developed standards, which include minimal information reporting requirements, terminologies, models and formats, are essential to structure the data in an unambiguous manner for humans and machines. Standards are more clearly defined and widely used for high dimensional data types where machine readable metadata are defined as part of the data capture (e.g. DICOM standards for imaging data) than for clinical data which then relates to the comparable FAIRness of resources for sharing OMICs and imaging versus clinical-epidemiological data. Five of the 47 clinical data registries used an eCRF that mapped to internationally accepted standards for clinical data, including International Classification of Diseases (ICD)-10 codes (n=2), Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT; n=1), Critical Care Data Dictionary (C2D2; n=1), Current Procedural Terminology (CPT) codes (n=1), Unified Medical Language System controlled unique identifier (UMLS CUI; n=1), and Anatomical Therapeutic Chemical (ATC) classification (n=1). Out of the platforms that share OMICs data, we only identified four that map to internationally accepted metadata or participant-level data standards. Out of these, two follow Minimum Information About a Microarray Experiment (MIAME) standards (Immport, GEO). Shared community-developed standards for participant-level data facilitate cross-resource analyses. Figure 2 shows the relationships between the 28 resources for sharing participant-level COVID-19 data included in the FAIRsharing Collection (https://fairsharing.org/collection/TDRCOVID19Participantleveldatasharingplatformsregistries), depicted as orange circles based on their implementation of community-developed standards, or shared eCRFs (as yellow shapes) or through explicit connections where participant-level data or metadata from one platform is recorded in another platform. The zoomed-in elements show the standards used by EGA (top) and by ENA database (bottom). In several instances, platforms for sharing high dimensional human or pathogen OMICs data or imaging data accept data in any community developed standard; Figure 2 includes a subset of the standards used in those cases.

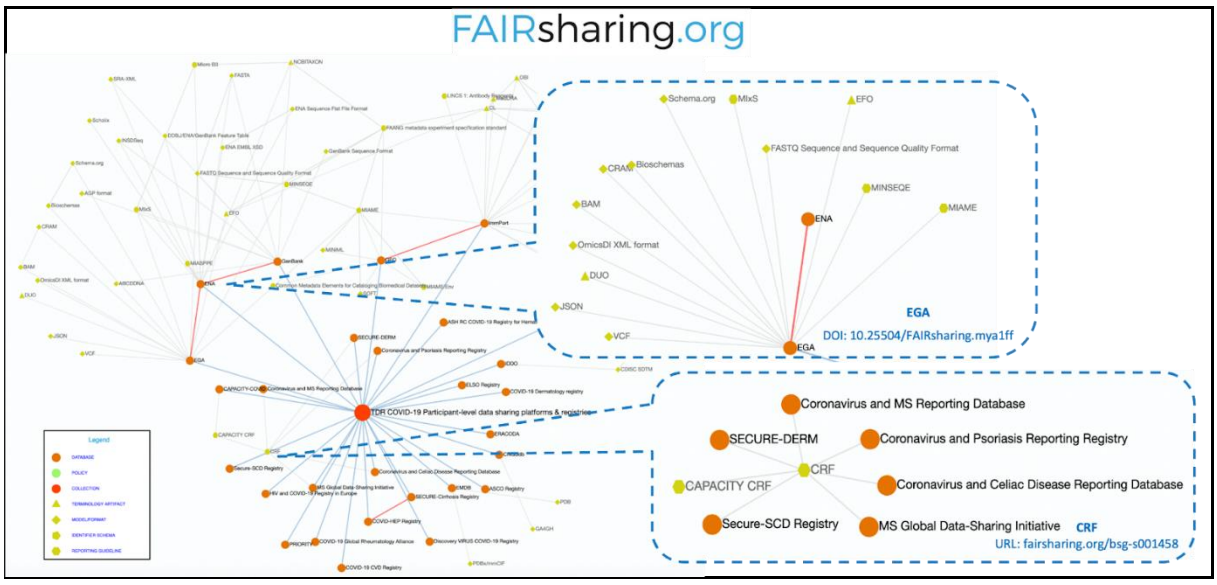


Fig 2. Interoperability of registries and platforms registered in the FAIRsharing.org COVID-19 Participant Level Data collection through shared use of community developed standards for participant-level data

While all but one platform (COVID-19 and MS) planned to share participant-level data, 18 of the 47 registries did not intend to share COVID-19-related, participant-level data. Four of the 24 registries and 3 of the 19 platforms that intended to share data did not provide any information on how to access data on their website. Additionally, 6 registries that intended to share data provided insufficient information on how to access data on their website (e.g. ACR CIRR, BADBIR, ENERGY, HOPE-2, COVI-PREG, and PsoProtect). Fourteen of the 24 registries (58%) and 10 of the 19 platforms (53%) that intended to share data did not have a data usage license or agreement mentioned on their website.

Seven platforms (33%) and four registries (8%) met all four of the criteria for FAIRness while four platforms (19%) and 13 registries (28%) met none of the criteria. Platforms that met all of the criteria for FAIRness were large, government-funded platforms that pre-existed COVID-19. The four registries that met all four criteria for FAIRness (ASCO, SCCM, LEOSS, ELSO) limited submission to member institutions. Registries and platforms that met none of the criteria for FAIRness were recently launched, COVID-19-specific resources that are still accepting data and may develop in terms of infrastructure and governance for COVID-19 data collection and sharing in the future. The application of a community-developed standard for participant-level data or metadata was the most commonly missed component of FAIRness, only two platforms and five registries that collected participant-level clinical data adopted a community-developed standard for that data, which is difficult to address retrospectively and limits the interoperability of COVID-19 data sharing efforts.

The protection of human subjects, the governance of and mechanism for data sharing, and engaging in meaningful benefit sharing with the research team that contributed data and the participants' source community are of central importance for ethical data sharing[2]. One platform (4%) and 18 registries (38%) did not plan to share participant-level data. For three of the 19 platforms (16%) and 4 of the 24 registries (17%) that had or intended to share data, we were not able to identify the data access mechanism through the website. Three of the 10 platforms that shared or intended to share human OMICs data were open access, two others had both open access and private data for which access was controlled by the data generators or a data access committee (DAC), one required DAC permission to access the data, two others required registration to access the data, and the other two did not have data access information on their website. Seventeen registries and five platforms that shared clinical data, five platforms that shared human OMICs data, and two platforms that shared linked clinical and human or pathogen OMICs or imaging data included a DAC to review data requests. Requests for data access for five platforms and one registry were decided by the data contributors.

Close to half of registries (N=20; 42%) stated that they were exempt from ethics review committee (ERC) oversight because they only collected de-identified data. While eight platforms only collected de-identified data, no platforms claimed ERC exemption. One platform and three registries stated that they would only accept participant-level data from groups that include broad consent for future use in their informed consent forms or have obtained waiver of consent.

Other than disseminating aggregate findings through a data dashboard (N=16 or 34% of registries; N=2; 10% of platforms), 19 platforms and 29 registries mentioned other forms of benefit sharing, including citation of the groups that provided data (9 platforms, 2 registries), citation of the data sharing resource itself (16 platforms, 9 registries), acknowledgement of data providers or co-

authorship on registry/platform-based publications (21 registries, 1 platform), or access to analytic tools (9 platforms, 1 registry). Only 9 registries, 4 platforms, 2 catalogue of platforms, and 1 federation of interoperable datasets and platforms mentioned any form of community engagement. Community engagement activities included: community forums to guide the overall direction of the platform (B1MG Stakeholder Portal; ICODA Public and Patient Voices Expert Group; PRIORITY Black, Indigenous, People of Color Health Equity and Birth Justice Core Group), involving patient representatives in the management of the registry (EULAR COVID-19 Registry, SECURE-AD, HS-COVID, COVID-19 and MS), work with health care providers to understand the implications for clinical practice (COVID-19 CVD Registry), and active engagement with researchers and their communities in LMIC through training and dissemination activities (IDDO).

There are no clearly defined metrics for determining whether a platform or registry is "successful." Data sharing resources reported the number of collaborating centers, datasets, participants represented by those datasets, and SARS-CoV-2 genome sequences, registered users and views or downloads of datasets to describe the breadth of data collection and dissemination work. Six registries and one platform (ReCoDID) did not include any information that could be used to characterize data submission or reuse. Two of those registries (ACS COVID-19 Registry, Transthoracic Echocardiography in COVID 19 Registry) did not specify if they would share data and the platform was not yet accepting data. Platforms and registries require significant initial and ongoing investments and the sustainability of data sharing resources is a major concern. Eleven registries and one platform received funding from more than one source. Eighteen platforms and one registry received government funding; two platforms and 23 registries received funding from related professional organizations or NGOs; and two platforms and seven registries received funding from industry sponsors. While some registries received funding from universities (N=12) and private donations (N=4), no platforms were funded by these sources.

Discussion

In this manuscript, we present the results of a year-long initiative by members of the COVID-19 Clinical Research Coalition to understand how participant-level data are being shared for COVID-19 response. In addition to monthly searches, we applied NLP to the CORD-19 database and consulted with colleagues that work on sharing human or pathogen OMICs data or clinical data in Europe, Canada, Africa, Latin America, and Asia to identify resources for collecting, harmonizing, and sharing, COVID-19-related participant-level data. We identified 68 platforms and registries for collecting, harmonizing, and sometimes sharing different types of COVID-19 data. For close to half of these, information that could be used to evaluate resource FAIRness or governance practices was not available on the website or in related documentation and was collected through our online survey. While we expect that these responses would still be current, existing resources have continued to evolve and additional data sharing registries or platforms that harmonize participant-level data have continued to develop since we completed our search in June 2021. Because the relevant data from registries are generally not machine readable, continuously updating and curating the results requires an important investment of time. A brief search conducted in October 2021 suggested that an additional 50 registries, including a number of vaccine registries, had been launched.

How do COVID-19 data sharing resources respond to existing data sharing principles?

While the importance of leveraging existing participant-level data and of connecting different data types at the participant level for COVID-19 response cannot be overstated, more resources for data sharing does not mean better data sharing. A number of groups developed principles for sharing different types of human research data prior to or during the COVID-19 pandemic. Below, we review

how COVID-19 data sharing platforms and registries map to the cross-framework principles of: collaboration; adherence to the FAIR principles; ethical issues, including transparent governance, protection of sensitive data, and community engagement; compliance with data protection laws; and evaluation of platform utility. The correspondence of data sharing resources to established principles is summarized in Table 2. Commonly shared challenges and recommendations for coordinated data sharing for COVID-19 response are presented in Table 3; stakeholder-specific recommendations are summarized in Table 4.

Collaboration

Data siloed by data type, comorbidities

The siloing of data by data type, comorbidity, and treatment increases the time required for sharing data when individuals with multiple comorbidities need to be entered into multiple databases and ultimately diminishes the utility of the data. The existing universe of disease-specific registries may lead to the exclusion of important populations affected by multiple comorbidities. Only a few of the 68 resources included clinical data that were linked to human and pathogen OMICs data at the participant level, which hinders efforts to respond to emerging and established VOCs.

WHO/IDDO/ISARIC eCRF & other efforts at facilitating prospective harmonization of participant-level COVID-19 data

In contrast to prior epidemics of emerging pathogens, the partnership between IDDO, ISARIC, and WHO, resulted in the rapid publication of a series of REDCap-based eCRFs that apply CDISC SDTM standards. Other than IDDO itself, only one of the 54 data sharing resources that collect COVID-19-related clinical data reported using the IDDO-ISARIC-WHO eCRFs, which represents a missed opportunity. There have been national efforts to facilitate the interoperability of COVID-19 data, including the US National Coordinator for Health Information Technology Logica COVID-19 Implementation Guide[19] which applies a Health Level 7 (HL7) Fast Healthcare Interoperability Resources (FHIR)-based library of COVID-19-related data elements and the UK NHS COVID-19 National Clinical Coding Standards.[20] Emerging international efforts, like the COVID-19 Interoperability Alliance,[21] which includes SNOMED/LOINC and RxNorm and addresses cross national interoperability of COVID-19 data, and the HL7 International Patient Summary Implementation Guide[22] and the European Health Data Space initiatives[23] have emerged to address cross national interoperability of electronic medical record (EMR) data.

Need for connections between research and clinical data streams

Selection bias, when the participants included in a study or database differ systematically from the population of interest, is an important consideration when accessing data uploaded to the platforms and registries described here. EMR data are an underutilized resource for surveillance and epidemic response[24,25] and represent a less selected population than the populations reflected in data that are manually entered data by hospital staff in disease-specific registries. Formidable barriers, including lack of interoperability, ethical concerns, and EMR-vendor or hospital-specific barriers to access[26], have prevented coordinated sharing of EMR data and likely led to the current universe of comorbidity- and population-specific registries. In addition to reducing the data entry burden incurred when data are shared through registries rather than EMR, there are compelling ethical arguments, including the duty of easy rescue, for using EMR data in the public health response to epidemics[27] and several ongoing initiatives to facilitate cross-national, interoperable EMR data[23,28-30].

FAIR principles and community standards

Our results show that platforms for sharing high dimensional participant-level health-related data (i.e. OMICS and imaging data) are better aligned with the FAIR principles than registries for sharing clinical or epidemiological data. This difference is explained in part because of the inclusion of machine-readable metadata and community-developed standards for participant-level data as part of the computational processing of high dimensional data, discipline-specific expectations regarding data availability and the use of community-developed standards, and limited regulatory oversight for observational health research. Registries for participant-level clinical data were less likely to be assessed for their adherence to the FAIR principles prior to the COVID-19 pandemic and how to measure the FAIRness of clinical or epidemiological data is an actively evolving conversation. The FAIR principles focus on machine readability and (re)use of data at scale; they do not address the quality and utility of the data resource and its content. The FAIR community continues work towards finalizing cross-disciplinary, cross-data type maturity indicators that can be implemented by any evaluation tool in order to yield consistent results and our evaluation of resource adherence to the FAIR principles should be read in the context of this evolving landscape. Funders can build on this initial evaluation of clinical research data sharing efforts by bringing together different stakeholders and disciplines to develop indicators to benchmark COVID-19 data sharing initiatives move towards FAIR data.

Seventeen of the 68 COVID-19-focused platforms and registries met none of the four basic criteria for FAIRness, which suggests a need to support those groups to enact basic steps to improve the platform or repository's adherence to the FAIR principles. The application of a community-developed standard for meta- or participant-level data is the most resource intensive and was the least commonly enacted of the four criteria. The role of data and metadata standards as essential elements for the consistent and meaningful reporting and sharing of information precedes the FAIR principles and their patchy implementation and use is a known issue[14,31]. Key challenges for interoperable clinical or epidemiological participant-level data and metadata include: (1) fragmentation with gaps and duplications and a lack of intra standard interoperability, which limits their consistent use, especially between medical and research areas; (2) differences in the governance and terms of use, especially between formal standard organizations and grass-root initiatives, which often limits contributions, extensions and modifications; and (3) lack of funds to implement the standards for participant-level data, train users, curate data, and support the standards life cycle, which is necessary to deal with the evolving technologies and emerging data types; and (4) a lack of standards for study metadata. In this analysis, we were not able to directly measure the uptake of community-developed standards by data resources and had to collect information on resource adoption of standards through an online survey. This snapshot of the standards landscape, which will continue to evolve on FAIRsharing.org, should facilitate conversations about the wider adoption of common standards and the need for cross-standard interoperability.

Cross-registry interoperability of participant-level data

The use of community-developed standards for participant level data and study metadata is an important precondition for interoperable data. The use of different community-developed standards for participant-level data is likely unavoidable and may be addressed retrospectively, as through the application of the Observational Medical Outcome Partnership common data model (OMOP CDM)[32]. That said, very few platforms or registries applied community-developed standards for participant-level data, further limiting the interoperability of these data sharing initiatives.

Comprehensive, machine-readable study and data sharing resource metadata are the first step toward interoperability. Funders may consider extending ongoing efforts to develop guidelines for user-defined metadata,[33] with a focus on clinical metadata, where, in contrast to OMICs and high dimensional imaging data, key metadata are not defined at data capture. Interoperability of

platform metadata and the application of shared standards for participant level data would represent important progress towards inter-platform or repository interoperability.

Ethical concerns & compliance with data protection laws

Ethical or governance related concerns must be addressed. There are several disparate frameworks for evaluating ethical concerns when sharing participant-level research and EMR data in the research response to a public health emergency. While there is general agreement that broad consent for future use should be sought when sharing de-identified EMR or research data[2], some groups argue that broad consent, and even informed consent, are not needed for sharing de-identified data[34,35]. Where broad consent for future use was not possible or sought, a waiver of consent may be granted for sharing participant-level data in keeping with the Council for International Organizations of Medical Sciences (CIOMS) guidance[36]. Most countries have legal frameworks for sharing participant level data in the public health response to an emergency, like the COVID-19 pandemic, irrespective of consent[37].

Most platforms and registries specified that they would only share de-identified data; seven of these platforms or registries indicated that they were exempt from ethical review because they were only sharing de-identified data. Maintaining data utility while preventing re-identification is an important challenge, especially in COVID-19 response where participant-level linkages between data types (i.e., pathogen and host OMICs data and clinical data) are important for detecting and responding to VOCs. Different definitions of what anonymized and pseudonymized data mean further complicate cross-initiative discussions and approaches[38]. Data sharing resources should consider establishing an independent ethics advisory committee, as distinct from a research ethics committee, that reflects community values and preferences for data sharing and can evaluate key ethical issues. Interoperable governance, consistent definitions, and common approaches to shared ethical and legal issues would both conserve scarce resources and facilitate explicit connections between related data sharing investments.

Equitable distribution of platforms

Multiple groups have highlighted the dangers of parachute research in the context of data sharing[39,40] and indicated that data sharing is perceived as widening existing disparities in access to funding and publication opportunities between researchers in high and low-and-middle income countries (LMIC)[41]. Platforms, in particular, represent long-term, significant investments in infrastructure and specialized expertise and the absence of data sharing platforms in LMIC represent a missed opportunity to support equitable, global data sharing for COVID-19 response.

Community engagement & benefit sharing

Resources that collect, harmonize, and share data have to be responsive to competing needs from a diversity of stakeholders, including data generating groups, research participants and their source communities, funders, end users, whether academic or commercial, the general public, and the Open Science Community. Community engagement is important for ethical data use and ensuring meaningful benefit sharing. When conducted properly, community engagement engenders trust, fosters understanding and ownership, and promotes the partnerships with communities that can support both data sharing and future research. The most frequently reported forms of benefit sharing were data dashboards and citation of the data contributing groups. Benefit sharing could also be in the form of documentation of data sharing-facilitated knowledge translation that could empower governments, the medical community, or the general public to take early action during a pandemic. Fewer than a quarter of registries and platforms reported engaging communities or investment in research capacity building.

Transparent governance

Data access models correspond to different political, ethical, administrative, regulatory, and legal contexts, resulting in different systems for the review and assessment of proposals to access the data. A common system to manage access involves review of an application to access the data by a centralized Data Access Committee (DAC). DACs review and evaluate proposals to access data and are central to ensuring that community values and preferences are reflected in data sharing decision making and setting public health priorities for data reuse. Independent commissions rather than individual researchers should be responsible for ensuring fair and equitable data sharing that balances the interests of data providers (e.g. publication), research participants or patients, and the open science and public health communities. Sixteen registries and five platforms that are or will share participant-level data included a DAC.

Several recent reviews explore best practice for DACs[3,6,42,43], which include, at a minimum, community representation, transparency and consistency regarding the process, criteria, and decisions around data requests and specific steps to avoid conflicts of interest between DAC members and dataset applicants. Further work is needed to define best practice for data governance with a focus on interoperable governance of data sharing efforts when responding to PHEICs. In public health emergencies, software approaches to shielded data access (e.g. DataSHIELD[44,45]), which allow for analysis without end users moving or "seeing" the data, may be a way to address ethical and legal concerns while ensuring timely data access for informed public health response.

Legal barriers to data sharing

Concerns about recent data protection laws, including GDPR, are likely correlated to siloed data and governance efforts, as when a platform deputizes individual institutions to manage data access rather than pooling responsibilities arising from data protection law, incl. establishing a centralized DAC to avoid distributed controllership. Lack of clarity in terminology[38] has contributed to inconsistent interpretations and applications of data protection laws within and beyond Europe which further hinder the interoperability of governance structures and initiatives that share interconnected data types. These fears have persisted in spite of provisions to support data sharing in the response to public health emergencies[46], including article 9(2)i of the GDPR which allows for the processing of sensitive personal data for reasons of public interest in the area of public health, including protection against serious cross-border threats to health, and Art. 49(1)d GDPR which provides an exemption for international data transfers if these are necessary for important reasons of public interest, which in practice became to include the public health response to infectious diseases.

Many countries lack national legal frameworks related to the cross-border transmission and transfer and sharing of participant-level, health-related data. As for the GDPR, its scope of application is broad and often results in the requirement for research entities in countries outside of Europe to comply with GDPR when interacting with EU-based institutions as when submitting, accessing, or receiving participant-level health data. The application of GDPR to the data processing activities of international organizations actively contributing to health research is contested. However, if EU-based organizations share data with international organizations, they must check the level of data protection within these organizations as this should be essentially equivalent with GDPR-level protection. Thus, besides the scope of application, transfer rules also quickly extend the reach of GDPR making it, on the practical level, the default data protection legislation. Additionally, collision rules are unclear when legal frameworks that prescribe data governance interact across national boundaries which leads to confusion regarding which rule to apply to the same data or a jointly conducted research activity and may further hinder data sharing.

Quantifying data resource utility

The public health imperative to share data to improve COVID-19 prevention and response has led to a proliferation of data sharing platforms and registries. There is a real need to understand the return on investment for these data sharing initiatives and to inform strategies to maximize the utility and sustainability of existing initiatives. While there have been a number of case studies that seek to demonstrate the utility of data sharing platforms, efforts to describe the public health-related benefits of sharing harmonized participant-level health-related data has been largely qualitative. Future research could identify markers for contributions to and usage of data sharing platforms and how the harmonization and dissemination of data facilitate research translation, build scientific networks, and lead to new fields of inquiry. In addition to understanding the utility of data sharing initiatives, clear metrics and quantitative approaches to assessing the downstream benefits and harms of data sharing could facilitate an exploration of ethical issues like whether data generated by researchers in LMIC benefit communities in LMIC and whether data contributors receive some measurable benefit in terms of novel funding applications, publications, collaborations, or research directions, from data sharing and producing the metadata needed to appropriately interpret that data.

Identifying and supporting successful investments

Platforms, and to a lesser extent, registries, require a significant investment of money and time. For example, the IDDO platform began with the World Wide Malaria Network in 2004 and an initial investment of over 20 million USD[47]. Investments in developing the governance and infrastructure for platforms that pre-existed the pandemic helped them transition rapidly to COVID-19 data collection. While established platforms, like IDDO, have shared data on close to 500K participants[48], COVID-19 platforms which were created during (e.g., CanCOGeN HostSeq and VirusSeq Portal) or slightly before the pandemic (ReCoDID) were not yet sharing data in July, 2021, when the platforms and registries overview dataset was finalized. Understanding which data sharing resources are "successful" in collecting and sharing data is as important as understanding how resources map to the FAIR principles and to best practice for ethical considerations related to international data sharing. We documented a number of metrics for evaluating the utility of data sharing resources, including the number of datasets, participants, genome sequences, and users. Future research should consider more nuanced measures of the impact of data sharing platforms and registries on preventing unnecessary research, improving the conduct of RCTs, and fast-tracking new discoveries or changes to clinical practice.

Coordinated data sharing for COVID-19 response

Collaboration between data sharing efforts, with a focus on the interoperability of related platforms based on interoperability of participant-level data and metadata through shared use of community developed standards, is perhaps the most important area for investment. The aggregation of standardized data across interoperable platforms or registries would help move towards the types of shared global analyses that could meaningfully inform the response to a global pandemic. The application of the same or interoperable standards for related study- and participant-level data is a necessary, but insufficient condition for inter-platform interoperability. In a few instances, connected platforms mean that data uploaded to one platform are reflected in another platform (e.g. SARS-CoV-2 OMICs data uploaded to the EMBL-EBI COVID-19 Data Portal or NCBI is included in INSDC), which enhances data findability and reuse. Large initiatives have emerged to connect platforms and registries within countries and regions, including the Health Data Research UK Innovation Gateway and the European COVID-19 Data Portal. Several initiatives exist to catalogue both COVID-19 data sharing initiatives and datasets (e.g. FAIRsharing; covid19dataindex). Coordination of COVID-19 clinical data sharing initiatives should include: the identification of several

core CDMs which can be meaningfully applied to research and EMR data, best practice for governance and addressing ethical and legal concerns, which can form the basis of an interoperable governance structure and common approach, where possible, to shared ethical and legal issues, and improved technical approaches for querying related data shared on disparate platforms or registries, including shielded approaches where participant-level data can be analyzed without being downloaded from the platform. Interoperability focused initiatives that focus on improving access to FAIR clinical and human and pathogen OMICs and high dimensional imaging data should be prioritized to facilitate the global response to VOCs.

Conclusion

Public health emergencies remind the public health and scientific communities of the urgent need to address unresolved barriers to sharing data in the context of infectious disease outbreaks. In contrast to the Zika and Ebola virus outbreaks, COVID-19 has ushered in a new era where researchers and funders need to shift their focus from supporting data sharing to promoting coordination between data sharing activities. The data sharing community, including funders, researchers, hospital networks, and public health authorities, need to move from a reactionary, fragmented response to a coordinated, synergistic approach. Ensuring that data sharing resources are as FAIR as possible and best practice for resource governance, transparency, community engagement, applicable legal frameworks, and recommended ethical (e.g. protection of research subjects, ERC review) and equitable practice (e.g. benefit sharing, community engagement) continues to be a key concern. In particular, interoperability within and between types (e.g., clinical, laboratory, OMICs) and sources (e.g. EMR, research study) of data should be a top priority for the current and future epidemics. Cloud-based platforms for data sharing represent a tremendous investment of financial resources and expertise. Clearly elaborated criteria for identifying successful platforms that apply best practice for governance and addressing ethical concerns, including benefit sharing, while meaningfully engaging the community can help funders focus investment by supporting good practice. While some duplication of effort should be expected, the current ecosystem of 47 registries and 21 platforms for sharing participant-level COVID-19 data that are not interoperable represents a lost opportunity and wasted resources. Given clear criteria for assessing platforms, funders, data generating groups, and the open science community can focus their efforts on a smaller number of well supported platforms and registries. Identifying the key political, ethical, administrative, regulatory, or legal motivations for the creation of disparate, non-interoperable, platforms for different diseases and datatypes is important for preventing continued investment in siloed data sharing efforts. Data sharing platforms generally have significant budgets because of the high cost of platform development and maintenance, retrospective data harmonization, and the governance of data sharing. All data sharing platforms were based in high income countries which raises questions of equity in the distribution of resources, concerns about the appropriate representation of the values and preferences of research teams and subjects based in LMIC, and in opportunities to build expertise in data curation and sharing. Data sharing is clearly on the policy agenda. We now need to move from fragmented, overlapping and competing data sharing efforts to a coordinated nexus of interconnected, longitudinal, participant-level data. Given the formidable barriers for such a cross-regional, cross-discipline initiative, we should start work now to be ready for the next global pandemic.

Acknowledgements

This work was funded by a TDR, the Special Programme for Research and Training in Tropical Diseases, grant to L.M. [P20-00007]. FAIRsharing is mainly funded by the Wellcome Trust, grant to S.A.S. [212930/Z/18/Z]. The authors would like to acknowledge Matias Bross for conducting the NLP analysis.

Author contributions

Conceptualization: L.M. Funding Acquisition: L.M.,R.T. Writing-original draft: L.M. Data collection: L.M., P.S., A.D. Data analysis: L.M., P.S., D.D., P.M. S.A.S. Writing—review & editing: L.M. P.S.,D.D., P.M.,R.T., F.M.G.,A.S., S.A.S.

Competing interests

No competing interests declared.

Table 1. Overview of platforms and registries for collecting, harmonizing, and sharing participant-level COVID-19-related data

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
1 Million Genomes & Beyond 1 Million Genomes Projects	ELIXIR Hub	ENG	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked)	Observ.	Both	No	Infected and non	Prespecified standard	N/A (Participant-level data not shared)
American College of Radiology COVID-19 Imaging Research Registry	American College of Radiology	US	No	Professional org.	Patients of all ages, with imaging examinations, in US	Clinical, imaging (linked)	Observ.	Longit.	No	Infected	Retrospective	DAC
American College of Surgeons COVID-19 Registry	American College of Surgeons	US	No	NS	Patients ≥18 yrs., admitted to the hospital, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	NS
American Society for Hematology Research Collaborative COVID-19 Registry for Hematologic Malignancy	American Society for Hematology Research Collaborative	US	No	Donation funded	Patients of all ages across the world with hematologic conditions	Clinical	Observ.	Cross-sect.	No	Infected	Prospective and retrospective	N/A (Participant-level data not shared)
American Society of Clinical Oncology Survey on COVID-19 in Oncology Registry	American Society of Clinical Oncology	US	No	Foundation funded	Patients of all ages with cancer, in US	Clinical	Observ.	Longit.	Yes	Infected	Prospective	DAC
British Association of Dermatologists Biologic and Immunomodulators Register	Uni. of Manchester	ENG	Yes	Industry	People of all ages, with psoriasis, in UK and IE	Clinical	Observ. and interventional	Longit.	No	Infected and non	Prospective	NS
Canadian COVID-19 Genomics Network - HostSeq Portal	Genome Canada	CA	No	Gov.	Patients of all ages, in CA	Clinical, human OMICs (linked)	NS	NS	No	Infected	Prospective and retrospective	DAC
Canadian COVID-19 Genomics Network - VirusSeq Data Portal	Genome Canada	CA	No	Gov.	Patients of all ages, in CA	Pathogen OMICs (unlinked)	Observ.	Cross-sect.	No	Infected	Prespecified standard	Open access

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
Cardiac Complications in Patients with SARS Corona virus 2 registry	Uni. Medical Center Utrecht	NL	No	NS	Patients ≥18 yrs., any location with cardiovascular complications	Clinical	Observ.	Longit.	No	Infected	Prospective	DAC
Center for International Blood and Marrow Transplant Research COVID-19 Data Collection	Center for International Blood and Marrow Transplant Research	US	Yes	Industry	Patients of all ages who are autologous and allogeneic hematopoietic cell transplantation recipients, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	DAC and resource management team
China National GeneBank DataBase	China National GeneBank	CN	Yes	NS	Patients of all ages, any location	Pathogen OMICs (unlinked)	Observ.	Cross-sect.	No	Infected	Prespecified standard	Public data is open access; Requests for controlled data overseen by reviewers/data owning organizations
Coronavirus and MS Reporting Database	Washington Uni.	US	No	Professional org.	Patients of all ages, with CNS demyelinating diseases, in North America	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
COVID-19 and MS – a global data sharing initiative	Multiple Sclerosis Data Alliance	BE	No	Professional org., industry	Patients of all ages, with MS, any location	Clinical	Observ. and interventional	Both	No	Infected	Prospective and retrospective	N/A (Participant-level data not shared)
COVID-19 CVD Registry	American Heart Association	US	No	Professional org., Donation funded	Patients ≥18 yrs. with cardiovascular complications, in US	Clinical	Observ.	Cross-sect.	No	Infected	Prospective and retrospective	DAC
COVID-19 Data Sharing/BR Initiative	Fundação de Amparo à Pesquisa do	BR	No	Uni., NGO	People of all ages who have undergone	Clinical	Observ.	Cross-sect.	No	Infected and non	No harmonization is work is	Open access

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
	Estado de São Paulo				COVID-19 testing, in Brazil						done. Data are uploaded with corresponding metadata	
COVID-19 Dermatology Registry	Massachusetts General Hospital	US	No	Professional org.	Patients of all ages, with dermatologic manifestations associated with or prior to COVID-19 infection, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC
COVID-19 Global Pediatric Rheumatology Database	Boston Children's Hospital	US	No	Professional org.	Pediatric patients ≤18 yrs.) with rheumatic disease, any location (except EU)	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC
COVID-19 Registry	Rice Uni.	US	No	Uni.	People ≥18 yrs., in US	Clinical	Observ.	Longit.	No	Infected and non	Prospective	Resource management team
COVID-Hepatology Registry	Translational Gastroenterology Unit, Uni. of Oxford	ENG	No	Gov., Uni., Professional org.	Patients of all ages, with liver disease or liver transplant, any location (except Americas, CN, JP, Koreas, and MN)	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
Discovery Viral Infection and Respiratory Illness Universal Study	Society of Critical Care Medicine	US	No	NGO	Patients of all ages, admitted to the hospital/ICU, any location	Clinical	Observ.	Longit.	Yes	Infected	Prospective	DAC
COVID-19 Registry Electron Microscopy Data Bank	EMBL-EBI	ENG	Yes	Gov.	People of all ages, any location	Imaging (unlinked)	Observ.	Cross-sect.	Yes	Infected and non	Prespecified standard	Open access

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
Electron Microscopy Public Image Archive	EMBL-EBI	ENG	Yes	Gov.	People of all ages, any location	Imaging (unlinked)	Observ.	Cross-sect.	Yes	Infected and non	Prespecified standard	Open access
European Academy of Neurology Neuro-covid Registry	European Academy of Neurology	AT	No	Professional org.	Patients ≥18 yrs., with neurological conditions, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	Resource management team
European Genome-Phenome Archive	EMBL-EBI	ENG	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked), pathogen OMICs (unlinked)	Observ. and interventional	Cross-sect.	No	Infected and non	Prespecified standard	DAC
European Nucleotide Archive	EMBL-EBI	ENG	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked), pathogen OMICs (unlinked)	Observ.	Cross-sect.	Yes	Infected and non	Prespecified standard	Open access to public data; data generator can restrict data access
Extracorporeal Life Support Organization Registry	Extracorporeal Life Support Organization	US	Yes	Professional org., Donation funded	Patients ≥16 yrs., on ECMO, any location	Clinical	Observ.	Longit.	Yes	Infected	Prospective	DAC
GenBank	NCBI-NLM	US	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked), pathogen OMICs (unlinked)	Observ. and interventional	Cross-sect.	No	Infected and non	Prespecified standard	Open access
Gene Expression Omnibus	NCBI-NLM	US	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked), pathogen OMICs (unlinked)	Observ. and interventional	Cross-sect.	Yes	Infected and non	Prespecified standard	Open access

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
GeneWeaver	The Jackson laboratory, Baylor Uni., The Uni. of Tennessee	US	Yes	Gov.	People of all ages, any location	Human OMICs (unlinked)	Observ. and interventional	Both	No	Infected and non	Prespecified standard	Open access to public data; data generator can restrict data access
Global Consortium to Study Neurological dysfunction in COVID-19 patients	Uni. of Pittsburgh	US	No	Professional org., Uni.	Patients of all ages, with neurological conditions, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	DAC
Global Hidradenitis Suppurativa COVID-19 Registry	Uni. of California San Francisco	US	No	NS	Patients of all ages, with hidradenitis suppurativa, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
Global Initiative on Sharing All Influenza Data	Freunde von GISAID e.V.	DE	Yes	Gov.	Patients of all ages, with influenza or corona inf., any location	Pathogen OMICs (unlinked)	Observ.	Cross-sect.	No	Infected	Prespecified standard	Resource management
Global Registry of COVID-19 in Pediatric Cancer	St. Jude Children's Research Hospital	US	No	Uni., Professional org.	Patients ≤18 yrs., with cancer, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	N/A (Participant-level data not shared)
Global Registry of COVID-19-related Diabetes	King's College London and Monash Uni.	ENG/AU	No	Uni.	Patients of all ages, with new-onset diabetes or acute complication of pre-existing diabetes, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	N/A (Participant-level data not shared)
Health Outcome Predictive Evaluation for COVID 19- 2	St Carlos Hospital	ES	No	NS	Patients of all ages, who have been discharged (deceased or alive) from any hospital center since Sept.	Clinical	Observ.	Longit.	No	Infected	Prospective	NS

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
					2020, any location							
Image Data Resource	Uni. of Dundee and Open Microscopy Environment	UK	Yes	Gov.	People of all ages, (geography NS)	Imaging (unlinked)	NS	Cross-sect.	Yes	Infected and non	Prespecified standard	Open access
Infectious Diseases Data Observatory	Oxford Uni.	ENG	Yes	Gov.	Patients of all ages, any location, with emerging pathogens or neglected diseases	Clinical	Observ. and interventional	Longit.	Yes	Infected	Prospective and retrospective	DAC
International COVID-19 and Pregnancy Registry	Centre Hospitalier Universitaire Vaudois	CH	No	Uni.	Pregn. women not considered minors, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	Resource management
International COVID-19 Data Alliance	Health Data Research UK	ENG	No	NGO	Patients of all ages, any location	Clinical	Observ. and interventional	Longit.	No	Infected	Prespecified standard	Resource management
Lean European Open Survey for SARS-CoV-2 Infected Patients	Uni. Hospital of Cologne and Goethe Uni. Frankfurt	DE	No	Professional org.	Patients of all ages, any location	Clinical	Observ.	Cross-sect.	Yes	Infected	Prospective	DAC and resource management
National Institute of Health - National COVID Cohort Collaborative	National Center for Data to Health by National Center for Advancing Translational Sciences hub sites	US	No	Gov.	Patients of all ages, in US	Clinical, human OMICs (linked), pathogen OMICs (linked), imaging (linked)	Observ.	Longit.	Yes	Infected	Prospective	DAC
Pediatric COVID-19 Case Registry	St. Jude Children's Research Hospital	US	No	NS	Patients <21 yrs. in US	Clinical	Observ.	Longit.	No	Infected	Prospective	DAC and resource management
Pregnancy Coronavirus Outcomes Registry	Uni. of California San Francisco Women's Health	US	No	Uni., Private donations	Pregn. or recently pregn.	Clinical	Observ.	Longit.	No	Infected	Prospective	N/A (Participant-

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
	Clinical Research Center				women ≥13 yrs., in US							level data not shared)
Psoriasis Patient Registry for Outcomes, Therapy and Epidemiology of COVID-19 Infection	Guy's and St Thomas' hospital	ENG	No	Professional org., Uni.	Patients of all ages, with psoriasis, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC
QMENTA imaging database	QMENTA	US	No	Industry	Patients of all ages, with imaging exam data, any location	Imaging (unlinked)	Observ.	Cross-sect.	No	Infected and non	Any format	Data generator
Reconciliation of Cohort data in Infectious Diseases	Heidelberg Uni. Hospital	DE	Yes	Gov.	Patients of all ages, any location	Clinical, human OMICs (linked)	Observ.	Longit.	No	Infected	Prospective and retrospective	DAC
SECURE-Celiac	Columbia Uni.	US	No	Industry	Patients of all ages, with celiac disease, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
SECURE-Liver	Uni. of North Carolina at Chapel Hill	US	No	Professional org.	Patients of all ages, with chronic liver disease or post-liver transplant, in North and South America, CN, JP, KR	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
SECURE-Psoriasis	Wake Forest School of Medicine	US	No	NS	Patients of all ages, with psoriasis, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
SECURE-Alopecia	National & International Skin Registry Solutions	IE	No	NGO	Patients of all ages, with alopecia, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
SECURE-Atopic Dermatitis	National & International Skin Registry Solutions	IE	No	NGO	Patients of all ages, with atopic dermatitis, any location	Clinical	Observ.	Cross-sect.	No	Infected and non	Prospective	N/A (Participant-level data not shared)
SECURE-Eosinophilic Esophagitis and Eosinophilic Gastrointestinal Diseases	Schneider Children's Medical Center in Israel	IL	No	NS	Patients of all ages, with eosinophilic gastrointestinal diseases, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	NS
SECURE-Inflammatory Bowel Disease	Uni. of North Carolina at Chapel Hill	US	No	Industry	Patients of all ages, with IBD, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC and resource management
SECURE-Sickle Cell Disease	Medical College of Wisconsin	US	No	NS	Patients of all ages, with sickle cell disease, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
SECURE-vascular anomalies	Children's Hospital of Philadelphia	US	No	NS	Patients of all ages, with vascular anomalies, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	N/A (Participant-level data not shared)
Society for Cardiovascular Magnetic Resonance COVID-19 Registry	Society for Cardiovascular Magnetic Resonance	US	Yes	NS	Patients of all ages, with cardiovascular complications and CMR data, any location	Clinical, imaging (linked)	Observ.	Longit.	Yes	Infected	Prospective	DAC and data generator
Society of Vascular and Interven. Neurology COVID-19 Registry	Cooper Uni. Hospital	US	No	Professional org.	Patients ≥18 yrs., with cerebrovascular complications, in US, ES, EG, R	Clinical	Observ.	Longit.	No	Infected	Prospective	Resource management
Surveillance of COVID-19 in Patients with T1D	T1D Exchange	US	No	Professional org., NGO, Industry	Patients of all ages, with type 1 diabetes, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	N/A (Participant-level data not shared)

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
The COVID-19 and Cancer Consortium	Vanderbilt Uni. Medical Center	US	No	NS	Patients ≥18 yrs., with cancer, in US, EU, AR, CA, MX, and UK	Clinical	Observ.	Longit.	Yes	Infected	Prospective	N/A (Participant-level data not shared)
The COVID-19 Global Rheumatology Alliance Registry	Uni. of California San Francisco	US	No	Industry	Patients >18 yrs., with rheumatic disease, any location (except EU)	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC and resource management
The database of Genotypes and Phenotypes	NCBI-NLM	US	Yes	Gov.	People of all ages, any location	Clinical, human OMICs (linked), pathogen OMICs (linked)	NS	Both	Yes	Infected and non	Prespecified standard	DAC
The European Alliance of Associations for Rheumatology COVID-19 Registry	European Alliance of Associations for Rheumatology	CH	No	Professional org.	Patients of all ages, with rheumatic disease, in EU	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	DAC and resource management
The European Renal Association COVID-19 Database	Uni. Medical Center Groningen	NL	No	Professional org., Industry	Patients ≥18 yrs., with kidney disease, in EU countries bordering Mediterranean	Clinical	Observ.	Longit.	No	Infected	Prospective	N/A (Participant-level data not shared)
The Immunology Database and Analysis Portal	Uni. of California San Francisco, Stanford Uni., Uni. of Buffalo, Technion - Israel Institute of Technology, and Northrop Grumman	US/IL	Yes	Gov.	Patients of all ages, any location	Human OMICs (unlinked), pathogen OMICs (unlinked), imaging (unlinked)	Observ. and interventional	Both	Yes	Infected	Prespecified standard	Data generator

Platform/ registry name	Lead org./ Consortium name	Location of coordinating center	Pre-existed COVID-19	Funding	Population-specific restrictions	Data types (linkage to clinical data)	Observ. or intervention data	Longit. or cross-sectional data	Community-developed standards for meta or participant-level data	COVID-19 status	Harmonization	Governance
The UK Coronavirus Cancer Monitoring Project	Uni. of Birmingham	ENG	No	Uni.	Patients of all ages, with cancer, in UK	Clinical	Observ.	Longit.	No	Infected	Prospective	NS
The UK Paediatric Oncology Coronavirus Cancer Monitoring Project	Uni. of Birmingham	ENG	No	Uni.	Patients <16 yrs., with cancer, in UK	Clinical	Observ.	Longit.	No	Infected	Prospective	NS
Thoracic Cancers International Covid 19 Collaboration Registry for Thoracic Cancers	Fondazione Istituto di Ricovero e Cura a Carattere Scientifico Istituto Nazionale Tumori, Vanderbilt Uni. Medical Center	IT/US	No	Professional org., Uni.	Patients of all ages, with thoracic cancer, any location	Clinical	Observ.	Longit.	No	Infected	Prospective	NS
Transthoracic Echocardiography in COVID 19 Registry	European Association of Cardiovascular Imaging	FR	No	Professional org.	Patients of all ages, with cardiovascular complications that undergo ECG, any location	Clinical	Observ.	Cross-sect.	No	Infected	Prospective	NS

AR, Argentina; AT, Austria; AU, Australia; BE, Belgium; BR, Brazil; CA, Canada; CH, Switzerland; CMR, cardiovascular magnetic resonance; CN, China; CNS, central nervous system; CRF, case report form; Cross-sect., cross-sectional; CVD, cardiovascular disease; DAC, data access committee; DE, Germany; DICOM, Digital Imaging and Communications in Medicine; EBI, European Bioinformatics Institute; ECG, echocardiographic examinations; ECMO, extracorporeal membrane oxygenation; EG, Egypt; EMBL, European Molecular Biology Laboratory; ENG, England; ES, Spain; EU, Europe; FR, France; Gov., government; IBD, inflammatory bowel disease; ICU, intensive care unit; IE, Ireland; IL, Israel; IT, Italy; JP, Japan; KR, Korea; Longit., longitudinal; MN, Mongolia; MS, multiple sclerosis; MX, Mexico; N/A, not applicable; NCBI, National Center for Biotechnology Information; NGO, non-governmental org.; NL, Netherlands; NLM, The National Library of Medicine; NS, not specified; Observ., observational; Pregn., pregnant; RO, Romania; SECURE, Surveillance Epidemiology of Coronavirus Under Research Exclusion; T1D, type 1 diabetes; UK, United Kingdom; Uni., university; US, United States of America.

Table 2. How COVID-19-related data sharing efforts map to established principles for data sharing

General domain	How COVID-19 data sharing resources correspond	7 GloPID-R Principles of Sharing Data in Public Health Emergencies[4]	COVID-19 NCS Data Sharing Principles[5]	International Code of Conduct for Data Sharing in Genomic Research[6]	GA4GH Framework for responsible sharing of genomic and health related data	CARE Principles for Indigenous Data Governance[7]
Collaboration	4 platforms and 2 registries had explicit connections between data types at the participant-level. Registry data were siloed by comorbidity, body system, and population		**X**	**X**	**X**	
FAIR data	7 platforms and 4 registries met the four basic criteria for FAIRness; 4 platforms and 13 registries met none of the criteria	**X**	**X**	**X**		
Ethical	1 platform and 3 registries were limited to data that included broad consent for future use or waiver of consent; 8 platforms and 21 registries only shared de-identified data; 2 platforms and 5 registries did not seek ERC approval; 2 registries and 9 platforms required or suggested citation of data providers in publications; 18 registries and 1 platform also mentioned acknowledgement of data providers in publications as a form of benefit sharing	**X**	**X**	**X**	**X**	**X**
Community engagement	4 platforms & 9 registries mentioned community engagement; 2 platforms and 16 registries included a data dashboard	**X**	**X**	**X**		**X**
Transparent governance	12 platforms & 1 registry were open access; 3 platforms & 17 registries had a DAC; Data providers decided data access for 5 platforms and 1 registry; 8 registries did not specify how data access was mediated	**X**	**X**	**X**	**X**	
Compliance with data protection laws	Not assessed		**X**	**X**	**X**	
Evaluate platform utility	20 platforms and 41 registries provided some measure of resource utility		**X**		**X**	**X**
Quality	Not assessed	**X**		**X**	**X**	
Timely	As of July 2021, participant-level data were available for 17 platforms &23 registries	**X**				

CARE, Collective Benefit, Authority to Control, Responsibility, Ethics; GA4GH, Global Alliance for Genomics and Health; GloPID-R, Global Health Security Initiative and Global Research Collaboration for Infectious Disease Preparedness; NCS, National Core Studies.

Table 3. Commonly shared challenges and recommendations for coordinated data sharing for COVID-19 response

Challenge	Recommendation
Interoperability & accessibility of EMR & sensitive research data	• Expand efforts to link EMR through shared standards (e.g. FHIR HL7 International Patient Summary; OMOP CDM) • Make EMR and sensitive research data (e.g. linked human OMICs and clinical data) accessible through shielded platform-based approaches to analyzing data without moving data (e.g. DataSHIELD) • Promote interoperability-based data reuse through promoting standards and providing open access code for commonly applied analyses (e.g., OHDSI)
Interoperability of EMR & research data	• Apply standards to observational research that are closely related to or the same as EMR data standards (e.g. SNOMED, LOINC) • Consider ethical imperative to use EMR data for improving health care
Resources siloed by data type, comorbidity, body system	• Address root causes of data silos (e.g., lack of interoperability or siloed data at the data generating group level, concerns about legal implications, vendor reluctance to share data, re-identification concerns)
Interoperability of platforms & registries	• Develop open access tools and guidance for meta harmonization across standards • Provide open access trainings on application of CDM-based approaches
Interoperability of governance structures	• Guidance on best practice for platform and registry governance • More sensitive methods of exploring possibility of re-identification

Resources have different degrees of FAIRness	• Register your resources in system like FAIRsharing to:
	  • become more discoverable
	  • indicate which data and metadata standards you implement
	  • describe your data accessibility mechanisms
	  • declare terms of use for your data
	• Provide support to users
	• Maximize connections with other resources

Benefit sharing & community engagement	• Develop guidance for best practice for community engagement • Provide support, foster accountability
Competition between data sharing resources	• Incentivize cooperation • Address technical barriers to inter-resource interoperability • Develop metrics for assessing the utility of data sharing platforms and registries • Develop guidance for best practice for platform governance & hold platforms and registries to those standards

CDM, common data model; DOI, digital object identifier; EMR, electronic medical records; FAIR, findable, accessible, interoperable, reusable; FHIR, Fast Healthcare Interoperability Resources; HL7, Health Level Seven International; LOINC, Logical Observation Identifiers Names and Codes; OHDSI, Observational Health Data Sciences and Informatics; OMOP, Observational Medical Outcomes Partnership; SNOMED, Systematized Nomenclature of Medicine.

Table 4. Recommended actions for stakeholders to support coordinated data sharing efforts for COVID-19 and beyond

Stakeholder	Recommendation
Funders	● Develop and implement metrics to quantify the return on investment in data sharing efforts ● Take concrete steps to make data more FAIR (e.g., recommend that resource register in system with machine readable metadata like FAIRsharing) ● Require prospective registration of observational studies in a repository that collects metadata and/or assigns a DOI ● Require a proportion of the budget to be put towards interoperability (e.g. community-developed standard for participant-level data) ● Require intervention and observational research studies to apply community-developed standards ● Support metacatalogues which facilitate data reuse by helping researchers obtain DOIs (e.g. FAIRsharing.org)
Journal editors	● Require DOI for participant-level dataset and research protocol to improve study and data discoverability ● Require implementation of a machine-readable FAIR checklist that covers issues related to data availability, interoperability, registration of metadata ● Incentivize data reuse
Regulators	● Create a regulatory body for observational research
Bioinformaticians, software developers, data stewards, and the open science community	● Conduct mixed methods research to understand where and when datasets are made available, including barriers and facilitators to using platforms with different governance structures ● Build connections between data sharing infrastructures (as when data uploaded to one platform is automatically available through other platforms) ● Expand open science initiatives to facilitate data reuse without data access (shielded approaches to data access, access through open source code and interoperable participant-level data and metadata) ● Implement (in tools, curation processes) and recommendations the use (in guidelines) of community-defined descriptive standards to enable structured reporting and meaningful reuse of data and metadata ● Refine and pilot specific indicators for evaluating the FAIRness of clinical and epidemiological data ● Foster compliance with best practice for governance related to future use of data or samples that is consistent with international ethics guidelines on the topic .through international ethics bodies
Legal	● Address real or perceived data protection law barriers (particularly re: GDPR) to data access through cross-national governance and legislation, and clarification of interpretation and application of existing laws ● Identify and address provincial/state/governorate-level legal barriers with regard to margins of implementation and interpretation ● Identify and address legal barriers related to reuse of data for various secondary purposes, including dependence of the primary purpose ● Identify and address legal barriers related to the reuse of data from protected minority groups under the perspectives of fairness and equity ● Identify misinterpretations of data protection roles (controller, processor, joint controller subprocessor) ● Clarify the connection between actors' data protection roles and their role in defining how data is used; work on modalities of involving data submitting communities, entities and actors into decisions about secondary data usage ● Work towards data protection governance that allows data subjects to assert their rights also in international data sharing contexts ● Clarify legal tools for international data transfers in emergency situations such as pandemics ● Define technical and data security measures necessary to protect international data transfers in emergency situations and if no established legal tool for the transfer has been defined, in order to offer data protection but also to allow data processing and interpretation ● Elaborate collision rules when legal frameworks interact across national boundaries
Ethics Advisory Bodies	● Raise awareness of health care providers, researchers, and other stakeholders about ethics guidelines for data sharing, data re-use, and re-use of medical data for research purposes ● Strengthen guidelines on privacy and confidentiality (and their limitations) within the scope of data re-use and data sharing. Focus on and support transparency and accountability

Stakeholder	Recommendation
	● Work across regulatory and legal entities and stakeholder groups to harmonize guidelines, and ensure consistency of approach in interpretation of shared ethical concerns ● Provide community-developed recommendations for community engagement related to different types of data sharing or data reuse-related infrastructures ● Provide community-developed recommendations on governance for different types of data sharing or data reuse-related infrastructures ● Require a section on FAIR data as part of ethics submissions for observational research
Data sharing platforms or registries	● Build expertise in related community-developed standards ● Meaningfully engage communities around data sharing ● Evaluate understanding of language around broad consent for future use

DOI, digital object identifier; EQUATOR, Enhancing the Quality and Transparency of Health Research; FAIR, findable, accessible, interoperable, reusable; GDPR, General Data Protection Regulation.

FAIR, ethical, and coordinated data sharing for COVID-19 response: a review of COVID-19 data sharing platforms and registries: Supplementary Information & References

Supplementary Table 1. Working definitions for resources for sharing participant-level data

Term	Definition	Approach to harmonization	Data types
Platform[49]	Combines big data tools and infrastructure. Major investment to continuously store, manage, mine big data sets (e.g. OMICs, imaging data).	Retrospective or prospective	May be limited to 1 data type or include various prespecified data types
Registry[50]	Collection of data physically stored in an assigned location. Low level of investment needed. Data generally entered or uploaded using the same case report form/data dictionary and focus on a particular disease, condition, or exposure.	Prospective	Generally limited to 1 specific data type
Dataverse[51]	Open source web application to share, preserve, cite, and explore research data of various types and with varying objectives.	Data is in its original form and not harmonized	Any
Datahub[52]	Data store that is an integration point for multiple datasets with different structures. Data are physically moved and stored together, however access permissions vary by data contributor.	Generally involves harmonization of data	Any
Data lake[52]	Central repository or pool of raw and untransformed data of any data type for an undefined purpose and requires other add-on tools to search or operationalize the data. Requires a low-level of investment.	Data is in its original form and not harmonized	Any
Data warehouse[53]	Data management tool that contains structured, filtered data that has already been processed and refined for a specific purpose allowing end users to perform further analytics.	No harmonization	Any
Data federation[54]	Technology wherein the data stored in different data sources are made accessible as one integrated virtual database and can be queried, transformed and accessed by data consumers. Data federation is a subset of data virtualization.	Data federation involves transformation, cleansing, and at times, the enrichment of data	Any
Data virtualization[52]	Data virtualization evolved from data federation with additional features and functionalities. According to different software developers, data virtualization has several capabilities beyond data federation including advanced security, query processing, and data transformation features.	Same as data federation	Any
Data catalogue[55]	Website with linkages to available datasets or platforms.	No harmonization	Does not host data

Supplementary Note 1. Google and Google Scholar Search Terms for COVID-19-related platforms and registries

(coronavirus OR COVID-19 OR :severe acute respiratory syndrome" OR coronavirus-2019 OR nCoV OR 2019nCoV OR 2019-novel CoV OR corona vir* OR coronavir* OR neocorona vir* OR neocoronavir* OR COVID OR COVID19 OR nCov 2019 OR nCov 19 OR SARS-CoV-2 OR SARS-CoV2 OR SARSCoV2 OR SARSCoV-2 SARS coronavirus 2 OR SARS-like coronavirus OR Severe Acute Respiratory Syndrome Coronavirus-2) AND (database oOR databases OR repository OR repositories OR registry OR registries OR platform OR platforms)

Supplementary Note 2. Natural Language Processing Strategy & Source Code

Natural language processing (NLP) was applied to the Covid-19 Open Research (CORD-19) Dataset[1] to identify additional COVID-19-related data sharing platforms and repositories. NLP was conducted in a Jupyter Notebook environment using R. We singled out titles in the CORD-19 database with desired relatability to the publication using the keywords: registry, registries, database, databases, platform, platforms, repository, repositories, IPD-MA, individual participant data meta-analysis, and data dashboard.

NLP was a useful approach to dealing with the CORD-19 resources in English as well as different languages because processing this data through automation is difficult to do without an understanding of the way humans speak and write naturally. By matching proper nouns and nouns to the root of the sentence, we identified citations in the CORD-19 database that stated the name of the database or registry. We matched appositional modifiers to the target search terms to pick up any missed items that the initial algorithm did not pick up, due to the format the title was written in its respective language.

The NLP R source code was as follows:

```
#Resources
#https://universaldependencies.org/en/dep/index.html
#https://universaldependencies.org/u/pos/
#https://nlp.stanford.edu/software/dependencies_manual.pdf

#code to single out titles with search terms
import csv
import os
import json
from collections import defaultdict
import spacy
cord_uid_to_text = defaultdict(list)
with open("metadata.csv", encoding="utf-8") as f_in:
reader = csv.DictReader(f_in)
for column in reader:
title = column['title']
abstract = column['abstract']
cord_uid = column['cord_uid']
title.lower()
if "registry" in title:
print(title + "\n")
elif "database" in title:
print(title + "\n")
elif "databases" in title:
print(title + "\n")
```

```
elif "registries" in title:
print(title + "\n")
elif "platform" in title:
print(title + "\n")
elif "platforms" in title:
print(title + "\n")
elif "repository" in title:
print(title + "\n")
elif "repositories" in title:
print(title + "\n")
elif "ipd-ma" in title:
print(title + "\n")
elif "data dashboard" in title:
print(title + "\n")
elif "individual participant data meta-analysis" in title:
print(title + "\n")

# with title : + "\n" + title + "\n"
#Code to match a proper noun (title of database) to the root of the
sentenc
import csv
import os
import json
from collections import defaultdict
import spacy
terms = ["registry", "registries", "database", "databases",
"platform", "pl
"ipd-ma", "data dashboard", "individual participant data meta-analy
cord_uid_to_text = defaultdict(list)
parser = spacy.load("en_core_web_lg")
with open("parsed_metadata.csv", encoding="utf-8") as f_in:
reader = csv.DictReader(f_in)
for column in reader:
title = column['title']
cord_uid = column['cord_uid']
pdoc = parser(title)
for token in pdoc:
if token.pos_ == "PROPN" and token.dep_ == "ROOT":
print(cord_uid)
else:
next

#code to match any appositional modifiers to specific target terms
import csv
import os
import json
from collections import defaultdict
import spacy
terms = ["registry", "registries", "database", "databases",
"platform", "pl
"ipd-ma", "data dashboard", "individual participant data meta-analy
cord_uid_to_text = defaultdict(list)
parser = spacy.load("en_core_web_lg")
with open("parsed_metadata.csv", encoding="utf-8") as f_in:
reader = csv.DictReader(f_in)
for column in reader:
```

```
title = column['title']
cord_uid = column['cord_uid']
pdoc = parser(title)
for token in pdoc:
if token.dep_ == "appos":
if token.orth_ in terms:
print(title)
```

Supplementary Note 3. REDCap Questionnaire for COVID-19 Data Sharing Resources

https://drive.google.com/file/d/15_R5kTimpNByDM07scBzX1ymOnfz-IU1/view?usp=sharing

Supplementary Note 4. Quantitative evaluation of adherence of platforms for sharing participant-level COVID-19 data to the FAIR Principles

We limited the quantitative evaluation to registries for participant-level clinical data because we could not apply the same metrics for resources for sharing different data types. For example, participant-level clinical registries are necessarily restricted access due to the sensitive nature of the data whereas databases for sharing pathogen OMICs data are open access. When comparing resources for sharing these data types, the difference in access to data does not mean that one resource is less accessible than the other. While discipline-specific FAIR criteria should be developed using a diverse panel of experts and stakeholders, we applied the aforementioned guidance to indicators used by the FAIRshake tool[10] algorithm to better align the tool's evaluation with the specific concerns that we thought would be most important to end users of registries of clinical data. None of the registries that collect and harmonize COVID-19 participant-level clinical data had been assigned a DOI prior to our review of the registries. Eighteen of the registries were assigned a DOI by FAIRsharing as part of our evaluation. Seventeen of the registries that we contacted to assign a DOI did not respond to these inquiries and we could not quantify their FAIRness.

Below, we review the criteria used to create the preliminary rubric for evaluating registries' for sharing participant-level clinical COVID-19 data adherence to the FAIR principles. These draft criteria will be presented to the Research Data Alliance, an international network of individuals and groups working to improve FAIR data. Blue text indicates metrics from the WHO team's Excel file. Green text is used for metrics from the FAIR Data Maturity Model Specification and Guidelines 2020.[12] Text with a strikethrough indicates text that was removed from the corresponding indicator in the source file.

Preliminary criteria for the application of the FAIR assessment rubric

Findable

- **PID (unique & persistent identifier) for the data (RDA-F1-01D / RDA-F1-02D)**
  Does the repository provide PIDs for the datasets therein?
  Value: Values will be the same for all the COVID-19 resources we assessed as it isn't clear (in a machine-actionable manner) what kind of PID they use, as we don't have data access. This means all the registries will fail this indicator.
- **Annotation with metadata (RDA-F2-01M)**

This is the only thing that could differ between resources: the quality / quantity of metadata annotation.
We can consider 3 levels:
i) nothing -> fail
ii) minimum (contact, description, ? to be defined) -> medium
iii) rich (? to be defined)
This could also be the sum of the criteria filled out in the WHO survey.
Value: Consider the metric as a success for every registry, as this was a criterion to enter them in FAIRsharing (a minimum set of metadata must be required to be inserted into FAIRsharing).

- **PID (unique & persistent identifier) for the metadata** **(RDA-F1-01M / RDA-F1-02M)**
  Does the metadata from the repository are assigned a unique & persistent identifier?
  Value: Values will be the same for all the COVID-19 resources because they have a PID for metadata **in FAIRsharing**.
- **Link between PID_data & PID_metadata** **(RDA-F3-01M)**
  Does the metadata include the unique & persistent identifier of the data?
  Value: this will always give a failure (PID_data: FAIL ; PID_metadata: FAIL) .
- **Findable on search engines** **(RDA-F4-01M)**
  Are the registries findable on search engines? (we can check if they are marked up with Schema.org
  Did not assess whether the registers were present on portals or institutional websites.
  Value: every registry gave a success except "European Renal Association COVID-19 Database" (when searching on Google, can't find https://www.eracoda.org/ link, but I can access thanks to FAIRsharing link or others websites that redirect to the link).

Accessible

- **Standard protocol and secured standard protocol (https, ftps)** **(RDA-A1-04M / RDA-A1.1-01M)**
  Is the metadata accessible via standard protocols such as HTTPS and FTPS?
  Value: Values are the same for all the registries, as they can be accessible by https website.
- **Authentication secure** **(RDA-A1.2-01D)**
  Is sensitive data accessible by secure authentication?
  Value: Consider REDCap secure (=success). For registries we don't know if there is an authentication (REDCap is not used), assigned a "Not Clear" value for these cases.
- **Metadata accessibility on the long term** **(RDA-A2-01M)**
  Will the metadata be accessible in the long term even if the resource disappears?
  Value: Values will be the same for all the COVID-19 resources because all the resources have a PID for metadata on FAIRsharing.
- **Contact information** **(no correspondence with RDA)**
  Is there any contact information available on the website (not sure that we should make a distinction between a "registry contact" and a PI contact: a registry contact is better for sustainability but there is a chance that these rapidly emerging resources will disappear just as quickly and, in this case, a PI contact is better).
  "Registry email"
  "PI email(s)"

"Registry contact email(s)"

Value: Marked as successful only if one of these criteria is met.

- **Contact information valid (no correspondence with RDA)**

  In the context of this type of repository, it is important that the contact responds. The WHO team sent a survey to the contact and they received or not an answer. I think we can only consider "responded to survey."

  "Responded to survey with detailed questions about data types, sharing, and governance"

  ~~"Notes from investigator on how to access data"~~

  Value: Marked as successful if "Responded to survey with detailed questions about data types, sharing, and governance" is met.

- **Data access (RDA-A1-01M)**

  Is there a clear description of the access to the data?

  "Link to description of how to access data"

  "Link to clearly specified governance mechanism for reviewing data access requests"

  ~~"Link to clearly specified criteria for reviewing data access requests"~~

  "Criteria for reviewing data access requests (from REDCap)"

  ~~"Who controls access to the data"~~

  Value: Ignored "Criteria for reviewing data access requests (from REDCap)." Averaged the other 4 criteria (green: 1 ; red: 0 ; yellow: 0,5).

  Removed the "who controls access to the data", it doesn't bring anything.

- **Data sharing (no correspondence with RDA)**

  Is the data shared? As raw data is not directly accessible, here we can assess if some summary / reports / data dashboard / scientific articles are available.

  "Data sharing status"

  ~~"Investigator explanation for why data won't be shared (write N/A if data will be shared)"~~

  "Is there a data dashboard / articles / reports available?"

  Value: Ignored "Investigator explanation for why data won't be shared (write N/A if data will be shared)". Marked as successful only if one of the two criteria is met.

Interoperable

- **Use of a controlled vocabulary (RDA-I1-01D)**

  Does the data use a knowledge representation expressed in a standardised format? It can be assumed here that the use of forms to insert patient data allows the use of a controlled vocabulary.

  "Link to COVID-19 CRF or data dictionary"

  Value: Success only if one of the two criteria is met.

- **Use of a FAIR controlled vocabulary (RDA-I2-01M / RDA-I2-01D)**

  Does the data use a knowledge representation expressed in a FAIR standardised format? We can remove OMICS standards and Imaging data standards because not appropriate for these registries.

  ~~"What formal standards does the platform apply for human OMICs data access?"~~

  ~~"Connection between CRF and existing standards (e.g., ICD 9-11, CDASH, SNOMED, LOINC)"~~

  "Uses ISARIC/WHO CRF (case report form)?"

  "Clinical-epidemiological standards used by registry"

  ~~"OMICs data shardards used by registry"~~

~~"Imaging data standards used by registry"~~

Value: no known standard could be identified in the registries (except for the Extracorporeal Life Support Organization Registry which uses a Clinical-epidemiological standard)

Removed OMICS and imaging standards as we only look at registries.

Removed the connexion between CRF and existing standard (removed from WHO spreadsheet + doublon with the use of clinical-epidemiological standards used by the registry).

- **Data contextualisation (related resources)** **(RDA-I3-01M)**

  Are there links to platforms in the same field to contextualise the register? Are there links to clinical trials on ClinicalTrials.gov ?

  "Links to related platforms"

  Value: success if yes, failure if no.

Reusable

- **Licence** **(RDA-R1.1-01M)**

  Is there a clear and accessible licence for re-use?

  "Data usage license"

  Value: success if yes, failure if no. Yes if found a link "terms of use", "terms of service", "copyright notice" on the corresponding website

- **Source of data** **(no correspondence with RDA)**

  Does metadata include provenance information ?

  "Who can enter data? (anyone, registered users of the platform, the platform hosts)"

  "How is data entered? (can data be uploaded? Is this through a REDCap data entry platform, etc?)"

  Value: I averaged the 2 criteria.

- **Use of community standard** **(RDA-R1.3-01M, RDA-R1.3-01D / RDA-R1.3-02M / RDA-R1.3-02D)**

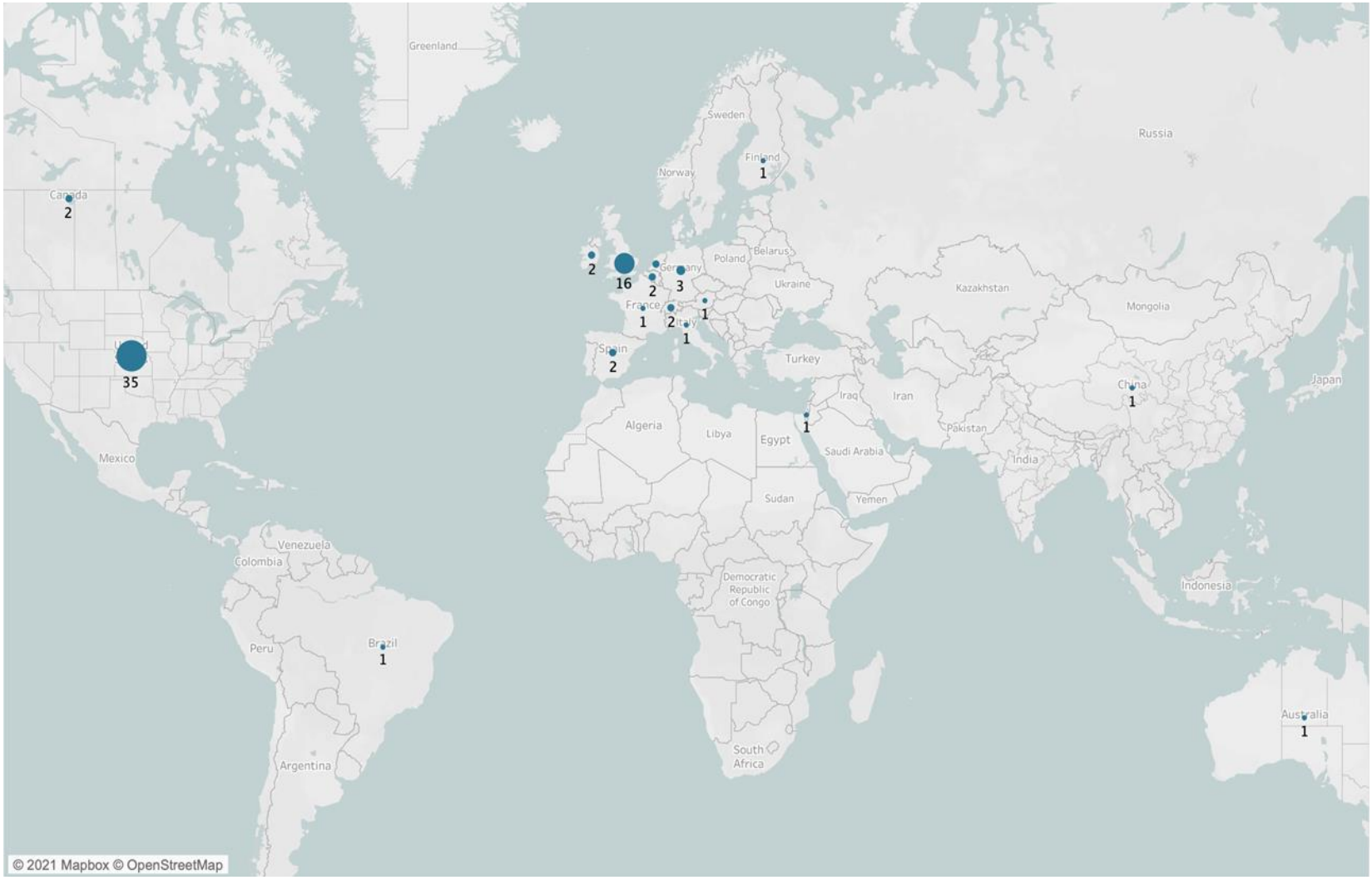
  Does data and metadata comply with a community standard ? Is data and metadata expressed in compliance with a machine-understandable community standard ?

  These are the results of: "Use of a FAIR controlled vocabulary". Almost all registries missed meeting this criterion.

  Value: Took the same results as the "Use of a FAIR controlled vocabulary" criteria (meaning failure for most of the registries). I just improved the score of the "except for the Extracorporeal Life Support Organization Registry" and "Discovery VIRUS COVID-19", as one standard is not sufficient to meet this criteria.

Supplementary Figure 1. Registry and platform-specific linkages between data types

Supplementary Figure 2. Summary distribution of platforms and registries for sharing participant-level COVID-19-related health data



*Circle size is proportional to the number of registries or platforms for collecting, harmonizing, sharing COVID-19

Supplementary Figure 3. Distribution of platforms and registries for sharing participant-level COVID-19-related health data

Blue circles represent the location of platforms. Orange circles represent the location of registries.

Supplementary Table 2. Potential data sharing resources identified through application of NLP to CORD-19 database

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
1	OutbreakTools: A new platform for disease outbreak analysis using the R software	Does not collect, harmonize, share COVID-19 participant-level data
2	Erratum: A biomimetic hybrid nanoplatform for encapsulation and precisely controlled delivery of theranostic agents	Does not collect, harmonize, share COVID-19 participant-level data
3	Corrigendum: A coral-on-a-chip microfluidic platform enabling live-imaging microscopy of reef-building corals	Does not collect, harmonize, share COVID-19 participant-level data
4	Project OPUS: Development and evaluation of an electronic platform for pain management education of medical undergraduates in resource-limited settings	Does not collect, harmonize, share COVID-19 participant-level data
5	A Self-Assembling Ferritin Nanoplatform for Designing Classical Swine Fever Vaccine: Elicitation of Potent Neutralizing Antibody	Does not collect, harmonize, share COVID-19 participant-level data
6	Amikacin pharmacokinetic/pharmacodynamic in intensive care unit: a prospective database	Does not collect, harmonize, share COVID-19 participant-level data
7	An Optimizing Multi-platform Source-to-source Compiler Framework for the NEURON MODeling Language	Does not collect, harmonize, share COVID-19 participant-level data
8	LitCovid: an open database of COVID-19 literature	Does not collect, harmonize, share COVID-19 participant-level data
9	DRDOCK: A drug repurposing platform integrating automated docking, simulations and a log-odds-based drug ranking scheme	Does not collect, harmonize, share COVID-19 participant-level data
10	Trialstreamer: a living, automatically updated database of clinical trial reports	Does not collect, harmonize, share COVID-19 participant-level data
11	REPP: A robust cross-platform solution for online sensorimotor synchronization experiments	Does not collect, harmonize, share COVID-19 participant-level data
12	SARS Grid--an AG-based disease management and collaborative platform	Does not collect, harmonize, share COVID-19 participant-level data
13	COVIDScholar: An automated COVID-19 research aggregation and analysis platform	Does not collect, harmonize, share COVID-19 participant-level data
14	LitCovid: an open database of COVID-19 literature	Duplicate (8)
15	DBCOVP: A database of coronavirus virulent glycoproteins	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
16	Telehealth Training During the COVID-19 Pandemic: A Feasibility Study of Large Group Multiplatform Telesimulation Training	Does not collect, harmonize, share COVID-19 participant-level data
17	Nanoplatforms for mRNA Therapeutics	Does not collect, harmonize, share COVID-19 participant-level data
18	Covid-19 Disease Simulation using GAMA platform	Does not collect, harmonize, share COVID-19 participant-level data
19	BioDynaMo: a general platform for scalable agent-based simulation	Does not collect, harmonize, share COVID-19 participant-level data
20	COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms	Does not collect, harmonize, share COVID-19 participant-level data
21	ERACODA: the European database collecting clinical information of patients on kidney replacement therapy with COVID-19	Registry that collects participant-level longitudinal clin-epi data about patients on kidney replacement therapy with COVID-19, harmonizes the data, but does not share the participant-level data
22	The National Gene Vector Biorepository's Pharm/Tox Database	Does not collect, harmonize, share COVID-19 participant-level data
23	MMDB: annotating protein sequences with Entrez's 3D-structure database	Does not collect, harmonize, share COVID-19 participant-level data
24	FSDB: A frameshift signal database	Does not collect, harmonize, share COVID-19 participant-level data
25	National Sample Vital Registration System: A sustainable platform for COVID-19 and other infectious diseases surveillance in low and middle-income countries	Does not collect, harmonize, share COVID-19 participant-level data
26	Virus-CKB: an integrated bioinformatics platform and analysis resource for COVID-19 research	Does not collect, harmonize, share COVID-19 participant-level data
27	Pathosphere.org: pathogen detection and characterization through a web-based, open source informatics platform	Does not collect, harmonize, share COVID-19 participant-level data (website is not available)
28	RNAcentral 2021: secondary structure integration, improved sequence search and new member databases	Does not collect, harmonize, share COVID-19 participant-level data
29	Infectome: A platform to trace infectious triggers of autoimmunity	Does not collect, harmonize, share COVID-19 participant-level data
30	CoVDB: a comprehensive database for comparative analysis of coronavirus genes and genomes	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
31	SARS-CoV-2 RECoVERY: A multi-platform open-source bioinformatic pipeline for the automatic construction and analysis of SARS-CoV-2 genomes from NGS sequencing data	Does not collect, harmonize, share COVID-19 participant-level data
32	The de.NBI / ELIXIR-DE training platform - Bioinformatics training in Germany and across Europe within ELIXIR	Does not collect, harmonize, share COVID-19 participant-level data
33	McQ - An open-source multiplexed SARS-CoV-2 quantification platform	Does not collect, harmonize, share COVID-19 participant-level data
34	Architectured Therapeutic and Diagnostic Nanoplatforms for Combating SARS-CoV-2: Role of Inorganic, Organic, and Radioactive Materials	Does not collect, harmonize, share COVID-19 participant-level data
35	Covid19db: An online database of trials of medicinal products to prevent or treat COVID-19, with a specific focus on drug repurposing	Does not collect, harmonize, share COVID-19 participant-level data
36	DatAC: A visual analytics platform to explore climate and air quality indicators associated with the COVID-19 pandemic in Spain	Does not collect, harmonize, share COVID-19 participant-level data
37	VIDA: a virus database system for the organization of animal virus genome open reading frames	Does not collect, harmonize, share COVID-19 participant-level data
38	MoonProt 3.0: an update of the moonlighting proteins database	Does not collect, harmonize, share COVID-19 participant-level data
39	CVTree update: a newly designed phylogenetic study platform using composition vectors and whole genomes	Does not collect, harmonize, share COVID-19 participant-level data
40	PhEVER: a database for the global exploration of virus–host evolutionary relationships	Does not collect, harmonize, share COVID-19 participant-level data
41	ELM—the database of eukaryotic linear motifs	Does not collect, harmonize, share COVID-19 participant-level data
42	IMG/VR: a database of cultured and uncultured DNA Viruses and retroviruses	Does not collect, harmonize, share COVID-19 participant-level data
43	SPRINT: a Cas13a-based platform for detection of small molecules	Does not collect, harmonize, share COVID-19 participant-level data
44	BiteOscope, an open platform to study mosquito biting behavior	Does not collect, harmonize, share COVID-19 participant-level data
45	Publisher Correction: Image Data Resource: a bioimage data integration and publication platform	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
46	Trialstreamer: A living, automatically updated database of clinical trial reports	Duplicate (10)
47	ThermoMutDB: a thermodynamic database for missense mutations	Does not collect, harmonize, share COVID-19 participant-level data
48	CEN-tools: an integrative platform to identify the contexts of essential genes	Does not collect, harmonize, share COVID-19 participant-level data
49	Early results of the Axium MicroFX for Endovascular Repair of IntraCranial Aneurysm (AMERICA) study: a multicenter prospective observational registry	Does not collect, harmonize, share COVID-19 participant-level data
50	CustusX: an open-source research platform for image-guided therapy	Does not collect, harmonize, share COVID-19 participant-level data
51	Neurologic manifestations in hospitalized patients with COVID-19: The ALBACOVID registry	Does not collect, harmonize, share COVID-19 participant-level data (closed study)
52	icumonitoring.ch: a platform for short-term forecasting of intensive care unit occupancy during the COVID-19 epidemic in Switzerland	Does not collect, harmonize, share COVID-19 participant-level data
53	[TCMATCOV--a bioinformatics platform to predict efficacy of TCM against COVID-19]	Does not collect, harmonize, share COVID-19 participant-level data
54	Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis	Does not collect, harmonize, share COVID-19 participant-level data
55	SECURE-Psoriasis: A de-identified registry of psoriasis patients diagnosed with COVID-19	Registry that collects participant-level cross-sectional clin-epi data about patients with psoriasis and COVID-19, harmonizes the data, but does not share the participant-level data
56	CMAUP: a database of collective molecular activities of useful plants	Does not collect, harmonize, share COVID-19 participant-level data
57	Commentary: The MISAGO registry: a rapid-exchange superficial femoral artery stent for a rapidly expanding field	Does not collect, harmonize, share COVID-19 participant-level data
58	PolarProtDb: a database of transmembrane and secreted proteins showing apical-basal polarity	Does not collect, harmonize, share COVID-19 participant-level data
59	SILK flow diverter for complex intracranial aneurysms: a Canadian registry	Does not collect, harmonize, share COVID-19 participant-level data
60	LncExpDB: an expression database of human long non-coding RNAs	Does not collect, harmonize, share COVID-19 participant-level data
61	Rapidemic, a versatile and label-free DNAzyme-based platform for visual nucleic acid detection	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
62	Donut PCR: a rapid, portable, multiplexed, and quantitative DNA detection platform with single-nucleotide specificity	Does not collect, harmonize, share COVID-19 participant-level data
63	DREIMT: a drug repositioning database and prioritization tool for immunomodulation	Does not collect, harmonize, share COVID-19 participant-level data
64	Comprehensive mapping of local and diaspora scientists: a database and analysis of 63951 Greek scientists	Does not collect, harmonize, share COVID-19 participant-level data
65	PFDB: a generic protein family database integrating the CATH domain structure database with sequence based protein family resources	Does not collect, harmonize, share COVID-19 participant-level data
66	COVID-19 management in heart transplanted recipients: registry of Almazov National Medical Research Centre	Does not collect, harmonize, share COVID-19 participant-level data (closed study)
67	Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability	Does not collect, harmonize, share COVID-19 participant-level data
68	WiFiMon: A mobility analytics platform for building occupancy monitoring and contact tracing using wifi sensing: Poster abstract	Does not collect, harmonize, share COVID-19 participant-level data
69	List N: Disinfectants for Use Against SARS-CoV-2 [database]	Does not collect, harmonize, share COVID-19 participant-level data
70	Neurologic manifestations associated with COVID-19: a multicentre registry	Registry that collects participant-level longitudinal clin-epi data about patients with neurological conditions and COVID-19, harmonizes the data, and shares this data. Has several prospective cohorts including adult and pediatric cohorts
71	Propedia: a database for protein-peptide identification based on a hybrid clustering algorithm	Does not collect, harmonize, share COVID-19 participant-level data
72	icumonitoring.ch: a platform for short-term forecasting of intensive care unit occupancy during the COVID-19 epidemic in Switzerland	Duplicate (52)
73	[TCMATCOV--a bioinformatics platform to predict efficacy of TCM against COVID-19]	Duplicate (53)
74	Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis	Duplicate (54)
75	Covigie, a platform for caregivers and care team coordinators	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
76	SECURE-Psoriasis: a de-identified registry of psoriasis patients diagnosed with COVID-19	Duplicate (55)
77	Global Hidradenitis Suppurativa COVID-19 Registry: a registry to inform data-driven management practices	Registry that collects participant-level cross-sectional clin-epi data about patients with hidradenitis suppurativa and COVID-19, harmonizes the data, but does not share the participant-level data
78	COVID-19 pandemic: Coroner's database of death inquiries with clinical epidemiology and total and excess mortality analyses in the District of Kildare March to June 2020	Does not collect, harmonize, share COVID-19 participant-level data
79	AgAcademy: a modal platform for scaling up e-learning in Indian agriculture in COVID times	Does not collect, harmonize, share COVID-19 participant-level data
80	CGAP: a new comprehensive platform for the comparative analysis of chloroplast genomes	Does not collect, harmonize, share COVID-19 participant-level data
81	METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive	Does not collect, harmonize, share COVID-19 participant-level data
82	ViPR: an open bioinformatics database and analysis resource for virology research	Does not collect, harmonize, share COVID-19 participant-level data
83	outbreaker2: a modular platform for outbreak reconstruction	Does not collect, harmonize, share COVID-19 participant-level data
84	MRPrimerV: a database of PCR primers for RNA virus detection	Does not collect, harmonize, share COVID-19 participant-level data
85	Anticovid, a comprehensive open-access real-time platform of registered clinical studies for COVID-19	Does not collect, harmonize, share COVID-19 participant-level data
86	The Brighton Collaboration standardized template for collection of key information for risk/benefit assessment of a Modified Vaccinia Ankara (MVA) vaccine platform	Does not collect, harmonize, share COVID-19 participant-level data
87	COVID-19 and its sequelae: a platform for optimal patient care, discovery and training	Does not collect, harmonize, share COVID-19 participant-level data
88	Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation	Does not collect, harmonize, share COVID-19 participant-level data
89	FragMAX: the fragment-screening platform at the MAX IV Laboratory	Does not collect, harmonize, share COVID-19 participant-level data
90	opvCRISPR: One-pot visual RT-LAMP-CRISPR platform for SARS-cov-2 detection	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
91	PURY: a database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures	Does not collect, harmonize, share COVID-19 participant-level data
92	HIT-COVID, a global database tracking public health interventions to COVID-19	Does not collect, harmonize, share COVID-19 participant-level data
93	A collection of designed peptides to target SARS-Cov-2 – ACE2 interaction: PepI-Covid19 database	Does not collect, harmonize, share COVID-19 participant-level data
94	SPDB: a specialized database and web-based analysis platform for swine pathogens	Does not collect, harmonize, share COVID-19 participant-level data
95	MarkerDB: an online database of molecular biomarkers	Does not collect, harmonize, share COVID-19 participant-level data
96	ROBOCOV: An affordable open-source robotic platform for COVID-19 testing by RT-qPCR	Does not collect, harmonize, share COVID-19 participant-level data
97	Neurological manifestations associated with COVID-19: a multicentric registry	Duplicate (70)
98	The spectrum of COVID-19-associated dermatologic manifestations: an international registry of 716 patients from 31 countries	Registry that collects participant-level cross-sectional clin-epi data about patients with dermatologic conditions and COVID-19, harmonizes the data, and shares this data
99	ADPriboDB 2.0: an updated database of ADP-ribosylated proteins	Does not collect, harmonize, share COVID-19 participant-level data
100	COVID-19 Disease Map, a computational knowledge repository of SARS-CoV-2 virus-host interaction mechanisms	Does not collect, harmonize, share COVID-19 participant-level data
101	RAPPID: a platform of ratiometric bioluminescent sensors for homogeneous immunoassays	Does not collect, harmonize, share COVID-19 participant-level data
102	DockCoV2: a drug database against SARS-CoV-2	Does not collect, harmonize, share COVID-19 participant-level data
103	Covid19Risk.ai: An open source repository and online calculator of prediction models for early diagnosis and prognosis of Covid-19	Does not collect, harmonize, share COVID-19 participant-level data
104	Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)	Does not collect, harmonize, share COVID-19 participant-level data
105	COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2	Does not collect, harmonize, share COVID-19 participant-level data

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
106	A2A: a platform for research in biomedical literature search	Does not collect, harmonize, share COVID-19 participant-level data
107	DPL: a comprehensive database on sequences, structures, sources and functions of peptide ligands	Does not collect, harmonize, share COVID-19 participant-level data
108	ViruSurf: an integrated database to investigate viral sequences	Does not collect, harmonize, share COVID-19 participant-level data
109	DescribePROT: database of amino acid-level protein structure and function predictions	Does not collect, harmonize, share COVID-19 participant-level data
110	CoV3D: a database of high resolution coronavirus protein structures	Does not collect, harmonize, share COVID-19 participant-level data
111	CORDITE: the curated CORona Drug InTERactions database for SARS-CoV-2	Does not collect, harmonize, share COVID-19 participant-level data
112	Pfam: The protein families database in 2021	Does not collect, harmonize, share COVID-19 participant-level data
113	Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm	Duplicate (71)
114	AlzGPS: a genome-wide positioning systems platform to catalyze multi-omics for Alzheimer's drug discovery	Does not collect, harmonize, share COVID-19 participant-level data
115	Swab-Seq: A high-throughput platform for massively scaled up SARS-CoV-2 testing	Does not collect, harmonize, share COVID-19 participant-level data
116	Engineering organoids: a promising platform to understand biology and treat diseases	Does not collect, harmonize, share COVID-19 participant-level data
117	Aging Atlas: a multi-omics database for aging biology	Does not collect, harmonize, share COVID-19 participant-level data
118	COVeAGE-DB: A database of age-structured COVID-19 cases and deaths	Does not collect, harmonize, share COVID-19 participant-level data
119	Validation of the Provincial Transfer Authorization Centre database: a comprehensive database containing records of all inter-facility patient transfers in the province of Ontario	Does not collect, harmonize, share COVID-19 participant-level data
120	COVID-19 Variants Database: A repository for Human SARS-CoV-2 Polymorphism Data	Tool that obtains data from the National Genomics Data Center (NGDC), Global Initiative on Sharing All Influenza Data (GISAID), and National Center for Biotechnology Information (NCBI) Genbank to help visualize the variants in the SARS-CoV-2 viral genome

No.	Title of citation for potential tool	Assessment of utility as resource for collecting, harmonizing and sharing COVID-19 participant-level data
121	OxCOVID19 Database: a multimodal data repository for better understanding the global impact of COVID-19	Does not collect, harmonize, share COVID-19 participant-level data
122	VIRsiRNAdb: a curated database of experimentally validated viral siRNA/shRNA	Does not collect, harmonize, share COVID-19 participant-level data
123	Guide to Immunopharmacology: a database to boost immunology education, research and therapy	Does not collect, harmonize, share COVID-19 participant-level data
124	GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences	Platform that obtains data from GISAID that allows users to browse, search and download single nucleotide variants at any individual or multiple SARS-CoV-2 genomic positions, or within a chosen genomic region or protein, or in a certain country/area of interest
125	DBatVir: the database of bat-associated viruses	Does not collect, harmonize, share COVID-19 participant-level data
126	H2V: a database of human genes and proteins that respond to SARS-CoV-2, SARS-CoV, and MERS-CoV infection	Does not collect, harmonize, share COVID-19 participant-level data
127	SARS2020: An integrated platform for identification of novel coronavirus by a consensus sequence-function model	Does not collect, harmonize, share COVID-19 participant-level data
128	The baculovirus expression vector system: A commercial manufacturing platform for viral vaccines and gene therapy vectors	Does not collect, harmonize, share COVID-19 participant-level data
129	CAPACITY-COVID: a European registry to determine the role of cardiovascular disease in the COVID-19 pandemic	Registry that collects participant-level longitudinal clin-epi data about patients with cardiovascular complications and COVID-19, harmonizes the data, and shares this data
130	Viral nanoparticles and virus-like particles: platforms for contemporary vaccine design	Does not collect, harmonize, share COVID-19 participant-level data
131	VirOligo: a database of virus-specific oligonucleotides	Does not collect, harmonize, share COVID-19 participant-level data
132	AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses	Does not collect, harmonize, share COVID-19 participant-level data

Supplementary Table 3. Application of FAIRshake[1] algorithm to registries for sharing COVID-19-related participant-level clinical data

	ASCO	COVID-19 CVD Registry	COVID-HEP Registry	COVID-19 Dermatology registry	Discovery VIRUS COVID-19	Extracorporeal Life Support Organization Registry	Pregnancy CoRonavIrus Outcomes RegIsTrY	COVID-19 Global Rheumatology Alliance	SECURE-Cirrhosis Registry	ASH Research Collaborative COVID-19 Registry for Hematologic Malignancy	Coronavirus and MS Reporting Database	HIV and COVID-19 Registry	Coronavirus and Celiac Disease Reporting Database	Coronavirus and Psoriasis Reporting Registry	SECURE-Sickle Cell Disease Registry	Coronavirus and Atopic Dermatitis (AD) / Alopecia Reporting Database	MS Global Data-Sharing Initiative	The European Renal Association COVID-19 Database
FINDABLE																		
PID (unique & persistent identifier) for the data																		
Annotation with metadata																		
PID (unique & persistent identifier) for the metadata																		
Link between PID_data & PID_metadata																		
Is the repository findable on search engine?																		
ACCESSIBLE																		
Standard protocol (http, ftp, smtp)																		
Secured standard protocol (https, ftps)																		
Authentication secure (Redcap is																		

	ASCO	COVID-19 CVD Registry	COVID-HEP Registry	COVID-19 Dermatology registry	Discovery VIRUS COVID-19	Extracorporeal Life Support Organization Registry	Pregnancy CoRonavIrus Outcomes RegIsTrY	COVID-19 Global Rheumatology Alliance	SECURE-Cirrhosis Registry	ASH Research Collaborative COVID-19 Registry for Hematologic Malignancy	Coronavirus and MS Reporting Database	HIV and COVID-19 Registry	Coronavirus and Celiac Disease Reporting Database	Coronavirus and Psoriasis Reporting Registry	SECURE-Sickle Cell Disease Registry	Coronavirus and Atopic Dermatitis (AD) / Alopecia Reporting Database	MS Global Data-Sharing Initiative	The European Renal Association COVID-19 Database
considered as secure)																		
Metadata accessibility long term																		
Contact information (registry or PI email)																		
Contact information valid																		
DATA ACCESS: Link to description of how to access data																		
DATA ACCESS: Link to clearly specified governance mechanism for reviewing data access requests																		
DATA ACCESS: Link to clearly specified criteria for reviewing data access requests																		

	ASCO	COVID-19 CVD Registry	COVID-HEP Registry	COVID-19 Dermatology registry	Discovery VIRUS COVID-19	Extracorporeal Life Support Organization Registry	Pregnancy CoRonavIrus Outcomes RegIsTrY	COVID-19 Global Rheumatology Alliance	SECURE-Cirrhosis Registry	ASH Research Collaborative COVID-19 Registry for Hematologic Malignancy	Coronavirus and MS Reporting Database	HIV and COVID-19 Registry	Coronavirus and Celiac Disease Reporting Database	Coronavirus and Psoriasis Reporting Registry	SECURE-Sickle Cell Disease Registry	Coronavirus and Atopic Dermatitis (AD) / Alopecia Reporting Database	MS Global Data-Sharing Initiative	The European Renal Association COVID-19 Database
DATA SHARING: Is/will the data be shared																		
DATA SHARING: Is there a data dashboard / articles / reports available ?																		
INTEROPERABLE																		
USE OF A CONTROLLED VOCABULARY: Link to COVID-19 CRF or data dictionary																		
USE OF A CONTROLLED VOCABULARY: Uses ISARIC/WHO CRF (case report form)?																		
USE OF A CONTROLLED VOCABULARY: Clinical-epidemiological standards used by registry																		

	ASCO	COVID-19 CVD Registry	COVID-HEP Registry	COVID-19 Dermatology registry	Discovery VIRUS COVID-19	Extracorporeal Life Support Organization Registry	Pregnancy CoRonavIrus Outcomes RegIsTrY	COVID-19 Global Rheumatology Alliance	SECURE-Cirrhosis Registry	ASH Research Collaborative COVID-19 Registry for Hematologic Malignancy	Coronavirus and MS Reporting Database	HIV and COVID-19 Registry	Coronavirus and Celiac Disease Reporting Database	Coronavirus and Psoriasis Reporting Registry	SECURE-Sickle Cell Disease Registry	Coronavirus and Atopic Dermatitis (AD) / Alopecia Reporting Database	MS Global Data-Sharing Initiative	The European Renal Association COVID-19 Database
Data contextualisation (related resources): Links to related platforms																		
REUSABLE																		
License																		
SOURCE OF DATA: Information about who can enter data (anyone, registered users of the platform, the platform hosts)																		
SOURCE OF DATA: Information about how is data entered (can data be uploaded? Is this through a REDCap data entry platform, etc?)																		

	ASCO	COVID-19 CVD Registry	COVID-HEP Registry	COVID-19 Dermatology registry	Discovery VIRUS COVID-19	Extracorporeal Life Support Organization Registry	Pregnancy CoRonavIrus Outcomes RegIsTrY	COVID-19 Global Rheumatology Alliance	SECURE-Cirrhosis Registry	ASH Research Collaborative COVID-19 Registry for Hematologic Malignancy	Coronavirus and MS Reporting Database	HIV and COVID-19 Registry	Coronavirus and Celiac Disease Reporting Database	Coronavirus and Psoriasis Reporting Registry	SECURE-Sickle Cell Disease Registry	Coronavirus and Atopic Dermatitis (AD) / Alopecia Reporting Database	MS Global Data-Sharing Initiative	The European Renal Association COVID-19 Database
Use of community standard																		

Green: the FAIR criteria are met; Red: the FAIR criteria are not met; Yellow: insufficient information

Supplementary Table 4. Summary of principles from health data sharing frameworks

General domain	7 GloPID-R Principles of Sharing Data in Public Health Emergencies[4]	COVID-19 NCS Data Sharing Principles[5]	International Code of Conduct for Data Sharing in Genomic Research[56]	GA4GH Framework for responsible sharing of genomic and health related data[6]	CARE Principles for Indigenous Data Governance[7]
Collaboration		Work collaboratively to actively share data to allow the scientific community to pool expertise, draw fresh insights, increase collective understanding.	Responsibility (Responsible governance should be shared between funders, generators and users of data. Investments in databases require coordination, strategy and long-term core funding. Mechanisms for building interoperability should be encouraged and appropriate management anticipated. Capacity building and recognition of all the data generators contributes to best practice.)	Education & training (Dedicate education and training resources so as to advance data sharing and data management and to constantly improve data quality and integrity.)	
FAIR data	Accessible (Data pertaining to PHEs should be shared with as few restrictions, either technical or legal, as possible. Providers of data should clearly indicate what, if any, conditions are in place, and for how long they apply.), FAIRness (The provision and use of data must be done in such a way that ensures fair treatment of all parties involved and recognition of their contributions. Further, any use of data should respect and acknowledge the provider and/or origin of the data and terms under which that data can be accessed and should reflect international commitments to benefits sharing.)	Ensure all data and associated code and tools generated through the studies are Findable, Accessible, Interoperable and Reusable (FAIR). Make research outputs, observations, code and tools generated from the studies open-source, rapidly and freely accessible as a public good.	Accessible (Facilitation of both the deposit of data and secure access to data are the foundations of data sharing. Curators of databases should promote sharing to generate maximum value. Harmonization of deposit, access procedures and use promotes accessibility, equity and transparency.)		
Ethical	Sharing of data must be done in accordance with applicable ethical and legal standards, ensuring beneficence and respect for	Consent (Ensure unconsented data is accessed through secure platforms accredited or working towards accreditation by the UK	Integrity (Mutual respect between all stakeholders is founded on personal and professional integrity. Prevention	Risk-Benefit Analysis (Consider the realistic harms and benefits of data sharing on and with individuals, families and communities, including	All data sharing should protect the privacy of individuals and the dignity of communities, while

General domain	7 GloPID-R Principles of Sharing Data in Public Health Emergencies[4]	COVID-19 NCS Data Sharing Principles[5]	International Code of Conduct for Data Sharing in Genomic Research[56]	GA4GH Framework for responsible sharing of genomic and health related data[6]	CARE Principles for Indigenous Data Governance[7]
	confidentiality, the privacy of individuals and the dignity of communities. This is essential for building the trust of the public and all stakeholders. Additional attention should be given to respect for, and alignment with, cultural norms.	Statistics Authority to comply standards established according to Digital Economy Act requirements, or working towards this status, which allows insights to be generated whilst maintaining privacy and data security)	of harms and anticipation of public concerns and scientific needs through foresight mechanisms encourage the development of common, prospective policies. Sanctions for breach of this Code or of other legal or ethical obligations must be clear.)	opportunity costs associated with both sharing and not sharing data. Conduct data sharing with a view towards minimizing harms and maximizing benefits to not just those who contribute their data, but also to society and health care systems as a whole.) Security (Establish proportionate data security measures that mitigate the risk of unauthorized access, data loss and misuse.)	simultaneously respecting the imperative to improve public health through the most productive use of data.
Community engagement	Equitable (data should be made available to all interested parties during a PHE at no cost, or at a cost recovery level only. This approach will help to ensure that all parties, including data providers and data users, have equal access to the data needed to collaborate and collectively deliver benefits to communities affected by a health emergency.)	Demonstrate active and ongoing engagement with patients and the public in the design, development and governance of their activities, to provide assurance that these activities are in the public interest.	Accountability (Inter-agency co-operation and funding fosters streamlined and efficient monitoring and good governance. Provisions should be made for ongoing public engagement that is tailored to the nature of the database and local cultures.)		Equitable (Any approach to the sharing of data should recognise and balance the needs of researchers who generate and use data, other analysts who might want to reuse those data, and communities and funders who expect health benefits to arise from research.)
Transparent governance	The process for sharing data and facilitating access should be clearly explained, outlining how and when the data can and cannot be shared and defining the associated descriptors of the data.	Be transparent in the use of personal data and respect the privacy and confidentiality of individuals, complying with legal requirements and ethical expectations at all times.	Key policies on publications, intellectual property, and industry involvement should be public. Websites that are accessible to the general public serve to provide feedback on progress and general results.	Develop clearly defined and accessible information on the purposes, processes, procedures and governance frameworks for data sharing.	
Compliance with data protection laws		Transparent use of personal data; respect the privacy and confidentiality of individuals (repeat of above)	Security (Trust and the promotion of data sharing rely on data management and security mechanisms and also on oversight of their functioning. Mechanisms for identifying and	Privacy, Data protection, Confidentiality (Comply with applicable privacy and data protection regulations at every stage of data sharing).	

General domain	7 GloPID-R Principles of Sharing Data in Public Health Emergencies[4]	COVID-19 NCS Data Sharing Principles[5]	International Code of Conduct for Data Sharing in Genomic Research[56]	GA4GH Framework for responsible sharing of genomic and health related data[6]	CARE Principles for Indigenous Data Governance[7]
			tracking data generators and users should be international.)		
Evaluate platform utility		Demonstrate value for money by using existing UK infrastructure and research investments as far as possible and using open competitions where necessary to develop new infrastructure capability.		Accountability (Put in place systems for data sharing that respect this Framework.Track the chain of data access and/or exchange to its source. Develop processes to identify and manage conflicts of interest. Implement mechanisms for handling complaints related to data misuse; for identifying, reporting and managing breaches; and for instituting appropriate sanctions.)	Efficient (Any approach to data sharing should improve the quality and value of research and increase its contribution to improving public health. Approaches should be proportionate and build on existing practice and reduce unnecessary duplication and competition.)
Quality	The minimum quality standard of data must be ensured by the provider while data users must also ensure that data processing, analysis and interpretation are conducted with an equal or greater application of quality standards. Appropriate and recognised data standards should be adhered to, while all relevant metadata, methodology, assumptions and experimental details should be provided with the data. This will ensure that any work conducted from the data takes into account the context in which the data was originally produced.		Irrespective of the discipline, scientists involved in data sharing should be *bona fide* researchers. Proof of academic or other recognized peer reviewed standing is essential. Harmonization of data collection and archiving methods and tools ensures validation of scientific quality. Collaboration promotes efficiency, sustainability and comparability.	Data quality & security (Store and process the data collected, used and transferred in a way that is accurate, verifiable, unbiased, proportionate, and current, so as to enhance their interoperability and replicability and also preserve their long-term searchability and integrity. Ensure feedback mechanisms on the utility, quality, security, and accuracy of data, and their annotations, with a view to improving quality and interoperability and appropriate re-use by others.)	
Timely	Timely				

CARE=Collective benefit, Authority to control, Responsibility, Ethics; GA4GH=Global Alliance for Genomics and Health; GloPID-R, Global Research Collaboration for Infectious Disease Preparedness; NCS=National Core Studies.

References

1 Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018, doi:10.1038/sdata.2016.18 (2016).

2 Bull, S., Roberts, N. & Parker, M. Views of Ethical Best Practices in Sharing Individual-Level Data From Medical and Public Health Research:A Systematic Scoping Review. *Journal of Empirical Research on Human Research Ethics* **10**, 225-238, doi:10.1177/1556264615594767 (2015).

3 Kalkman, S., Mostert, M., Gerlinger, C., van Delden, J. J. M. & van Thiel, G. J. M. W. Responsible data sharing in international health research: a systematic review of principles and norms. *BMC Medical Ethics* **20**, 21, doi:10.1186/s12910-019-0359-9 (2019).

4 Group, G.-R. D. S. W. Principles of data sharing in public health emergencies. (Global Research Collaboration for Infectious Disease Preparedness (GloPID-R) 2018).

5 *COVID-19 National Core Studies Data Sharing Principles*, <https://www.hdruk.org/covid-19/covid-19-national-core-studies/ncs-data-sharing-principles/> (2021).

6 Knoppers, B. M. Framework for responsible sharing of genomic and health-related data. *HUGO J* **8**, doi:10.1186/s11568-014-0003-1 (2014).

7 Carroll, S. R. *et al.* The CARE Principles for Indigenous Data Governance. *Data Science Journal* **18**, 43, doi:http://doi.org/10.5334/dsj-2020-043 (2020).

8 Lu Wang, L. *et al.* CORD-19: The Covid-19 Open Research Dataset. *ArXiv*, arXiv:2004.10706v10702 (2020).

9 Harris, P. A. *et al.* Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of biomedical informatics* **42**, 377-381, doi:10.1016/j.jbi.2008.08.010 (2009).

10 Clarke, D. J. B. *et al.* FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources. *Cell Systems* **9**, 417-421, doi:https://doi.org/10.1016/j.cels.2019.09.011 (2019).

11 Wilkinson, M. D. *et al.* Evaluating FAIR maturity through a scalable, automated, community-governed framework. **6**, 1-12 (2019).

12 FAIR Data Maturity Model. Specification and Guidelines (1.0). (2020).

13 Rosnet, T. & Gaignard, A. *FAIR-Checker: How FAIR is my resource?*, <https://fair-checker.france-bioinformatique.fr/base_metrics> (2021).

14 Sansone, S.-A. *et al.* FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology* **37**, 358-367, doi:10.1038/s41587-019-0080-8 (2019).

15 Nalbandian, A. *et al.* Post-acute COVID-19 syndrome. *Nature Medicine* **27**, 601-615, doi:10.1038/s41591-021-01283-z (2021).

16 Garg, M. *et al.* The Conundrum of 'Long-COVID-19': A Narrative Review. *Dovepress* **14**, 2491-2506, doi:https://doi.org/10.2147/IJGM.S316708 (2021).

17 Munblit, D. *et al.* Legacy of COVID-19 infection in children: long-COVID will have a lifelong health/economic impact. *Archives of Disease in Childhood*, archdischild-2021-321882, doi:10.1136/archdischild-2021-321882 %J Archives of Disease in Childhood (2021).

18 Miller, F. *et al.* Prevalence of persistent symptoms in children during the COVID-19 pandemic: evidence from a household cohort study in England and Wales. 2021.2005.2028.21257602, doi:10.1101/2021.05.28.21257602 %J medRxiv (2021).

19 *Logica COVID-19 FHIR Profile Library IG. 0.13.1 - ci-build*, <https://covid-19-ig.logicahealth.org/index.html> (2021).

20 *COVID-19 National Clinical Coding Standards and Guidance*, <https://hscic.kahootz.com/connect.ti/t_c_home/view?objectID=19165328> (2020).

21 *COVID-19 Interoperability Alliance*, <https://covid19ia.org/about/> (

22 Cangioli, G., Hausam, R. & Macary, F. *International Patient Summary Implementation Guide. 1.0.0 - STU 1*, <http://hl7.org/fhir/uv/ips/> (2020).

23 Vayena, E. Value from health data: European opportunity to catalyse progress in digital health. *The Lancet* **397**, 652-653, doi:10.1016/S0140-6736(21)00203-8 (2021).

24 Madhavan, S. *et al.* Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers. *Journal of the American Medical Informatics Association* **28**, 393-401, doi:10.1093/jamia/ocaa287 %J Journal of the American Medical Informatics Association (2020).
25 Dagliati, A., Malovini, A., Tibollo, V. & Bellazzi, R. Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Briefings in Bioinformatics* **22**, 812-822, doi:10.1093/bib/bbaa418 %J Briefings in Bioinformatics (2021).
26 Kuo, A. M.-S., Thavalathil, B., Elwyn, G., Nemeth, Z. & Dang, S. The Promise of Electronic Health Records to Promote Shared Decision Making: A Narrative Review and a Look Ahead. *Medical Decision Making* **0**, 0272989X18796223, doi:10.1177/0272989x18796223 (2018).
27 Porsdam Mann, S., Savulescu, J. & Sahakian, B. J. Facilitating the ethical use of health data for the benefit of society: electronic health records, consent and the duty of easy rescue. **374**, 20160130, doi:doi:10.1098/rsta.2016.0130 (2016).
28 Johnson, A. E. W., Stone, D. J., Celi, L. A. & Pollard, T. J. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association* **25**, 32-39, doi:10.1093/jamia/ocx084 (2018).
29 Lee, H.-A. *et al.* Global Infectious Disease Surveillance and Case Tracking System for COVID-19: Development Study. *JMIR Med Inform* **8**, e20567, doi:10.2196/20567 (2020).
30 Rogers, P., Wang, D. & Lu, Z. Medical Information Mart for Intensive Care: A Foundation for the Fusion of Artificial Intelligence and Real-World Data. **4**, doi:10.3389/frai.2021.691626 (2021).
31 Sansone, S.-A. & Rocca-Serra, P. Interoperability Standards-Digital Objects in Their Own Right. m9 (Wellcome Trust, London, England, 2016).
32 Voss, E. A. *et al.* Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association* **22**, 553-564, doi:10.1093/jamia/ocu023 %J Journal of the American Medical Informatics Association (2015).
33 Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J Am Med Inform Assoc* **22**, 1148-1152, doi:10.1093/jamia/ocv048 (2015).
34 Pearce, N. & Smith, A. H. Data sharing: not as simple as it seems. *Environmental Health* **10**, 107, doi:10.1186/1476-069X-10-107 (2011).
35 de Wolf, V. A., Sieber, J. E., Steel, P. M. & Zarate, A. O. Part I: What Is the Requirement for Data Sharing? *IRB: Ethics & Human Research* **27**, 12-16, doi:10.2307/3563537 (2005).
36 (CIOMS), C. f. I. O. o. M. S. International ethical guidelines for health-related research involving humans. Report No. 9290360887, (Geneva, Switzerland, 2016).
37 Phillips, M. International data-sharing norms: from the OECD to the General Data Protection Regulation (GDPR). *Human Genetics* **137**, 575-582, doi:10.1007/s00439-018-1919-7 (2018).
38 Ohmann, C. *et al.* Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* **7**, doi:10.1136/bmjopen-2017-018647 (2017).
39 Serwadda, D., Ndebele, P., Grabowski, M. K., Bajunirwe, F. & Wanyenze, R. K. Open data sharing and the Global South—Who benefits? *Science* **359**, 642-643, doi:10.1126/science.aap8395 (2018).
40 Conton, B. Build the Ebola database in Africa. *Nature* **551**, 143, doi:https://doi.org/10.1038/d41586-017-05676-4 (2017).
41 Bull, S. *et al.* Best practices for ethical sharing of individual-level health research data from low- and middle-income settings. *J Empir Res Hum Res Ethics.* **10**, doi:10.1177/1556264615594606 (2015).
42 Meyer, M. N. Practical Tips for Ethical Data Sharing. *Advances in Methods and Practices in Psychological Science* **1**, 131-144, doi:10.1177/2515245917747656 (2018).

43 Shabani, M., Knoppers, B. M. & Borry, P. From the principles of genomic data sharing to the practices of data access committees. *EMBO Molecular Medicine* **7**, 507-509, doi:https://doi.org/10.15252/emmm.201405002 (2015).
44 Gaye, A. *et al.* DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology* **43**, 1929-1944, doi:10.1093/ije/dyu188 (2014).
45 Blacketer, C. in *The Book of OHDSI* Ch. 4, (The OHDSI Community, 2021).
46 McLennan, S., Celi, L. A. & Buyx, A. COVID-19: putting the General Data Protection Regulation to the test. *JMIR Public Health Surveill* **6**, e19279, doi:10.2196/19279 (2020).
47 Pisani, E. & Botchway, S. Sharing individual patient and parasite-level data through the WorldWide Antimalarial Resistance Network platform: A qualitative case study. *Wellcome open research* **2**, 63-63, doi:10.12688/wellcomeopenres.12259.1 (2017).
48 *What is the IDDO Cognitive City and how do I use it?*, <https://iddo.cognitive.city/cognitive/welcome> (2021).
49 Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *Journal of Big Data* **6**, 54, doi:10.1186/s40537-019-0217-0 (2019).
50 Gliklich, R. & Dreyer, N. in *Registries for Evaluating Patient Outcomes: A User's Guide* (eds Gliklich RE & Dreyer NA) (Agency for Healthcare Research and Quality, 2010).
51 King, G. *A Comparative Review of Various Data Repositories*, <https://dataverse.org/blog/comparative-review-various-data-repositories> (2017).
52 *Data Hub vs Data Lake vs Data Virtualization*, <https://www.marklogic.com/product/comparisons/data-hub-vs-data-lake/> (2021).
53 *What Is a Data Warehouse?*, <https://www.oracle.com/sg/database/what-is-a-data-warehouse/> (2021).
54 Allemang, D. & Hendler, J. in *Semantic Web for the Working Ontologist (Second Edition)* (eds Dean Allemang & Jim Hendler) 51-60 (Morgan Kaufmann, 2011).
55 *Data Catalog*, <https://old.nnlm.gov/data/thesaurus/data-catalog> (
56 Knoppers, B. M. *et al.* Towards a data sharing Code of Conduct for international genomic research. *Genome Medicine* **3**, 46, doi:10.1186/gm262 (2011).