Data sharing for COVID-19 Landscape analysis

Nicola Perrin, Giulia Tomba

October 2020

Contents

- Overview

- Methodology

- Data portals and platforms

- Identifying data sources
  - Trial trackers and registries
  - COVID-19 datasets

- COVID-19 initiatives
  - Research consortiums and collaborations
  - Data science initiatives
  - Data sharing policy initiatives

- Next steps

Overview

- There has been a huge investment in research to tackle COVID-19 in the last six months. According to the UKCDR COVID tracker, 1,858 studies had been funded, mobilized or repurposed in July 2020; by November that figure had increased to 5,046 studies across 102 countries.

- The number of COVID-related clinical trials rose from 500 in April 2020 to 2,232 by September 2020. An ever-increasing amount of relevant data is therefore being generated and collected, but it is not always easy to know how to access it.

- A large number (200+) of COVID-19 data-related collaborations and initiatives have also sprung up. These range from small-scale specific projects to initiatives led by well-established programmes.

- With such a complex and busy landscape, it is often difficult to keep up-to-date with what is happening, and to be able to distinguish between promise and action. The lack of coordination across different initiatives poses a risk of duplication.

- ICODA conducted a mapping exercise to build our understanding of COVID data-related activities. We are sharing the initial findings from this internal work to help inform our partners and reduce duplication. We intend to collaborate with others to keep this information up-to-date and widely available.

PLEASE NOTE: the COVID-19 data landscape is evolving rapidly, and this mapping is not exhaustive

Methodology

- Information was retrieved through desk research on sources available on the Internet (including relevant stakeholders' webpages, press releases, specialised forums, scientific publications). We also learnt from other mapping exercises, including those by UKCDR, CAIC, NICE and EBI.

- Most of the data collection took place in July 2020, with a further update in September 2020.

- The main focus of the mapping was on data collections that were directly linked to COVID-related data, e.g. dashboards, databases, data repositories, datasets, libraries of datasets, platforms/portals for accessing datasets, registries of research or datasets (disease-specific registries, research trackers). The aim was to capture information about the types of data involved and the geographical scope.

- The mapping also captured initiatives related to COVID that might be collecting data, and other activities relating to data sharing policy initiatives, data science collaborations, and work to harmonise data standards.

- We are grateful to Giulia Tomba who undertook the bulk of the work.

Limitations
- In several cases, attribution to a specific category is not unequivocal. In almost all cases, it was based on the definition provided by the organisers/stakeholders themselves. When this was not available, the closest match was chosen based on comparison with similar activities.
- Some aspects would benefit from direct communication with relevant organisations: e.g., the actual stage of development and impact of a large number of initiatives is unclear
- The research work was mainly carried out in English. This may have prevented access to material available in other languages (mainly, initiatives in Latin America and Asia).

Key finding

The landscape of COVID-19 data-related activities is busy and complex. Key challenges include fragmentation, duplication and a lack of harmonisation. There is a real need for collaboration and coordination.
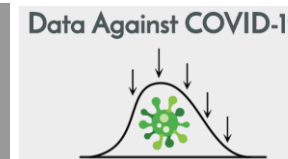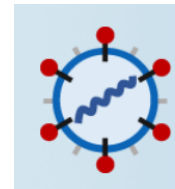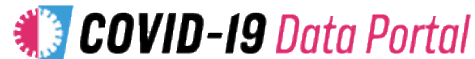
Our mapping considered:

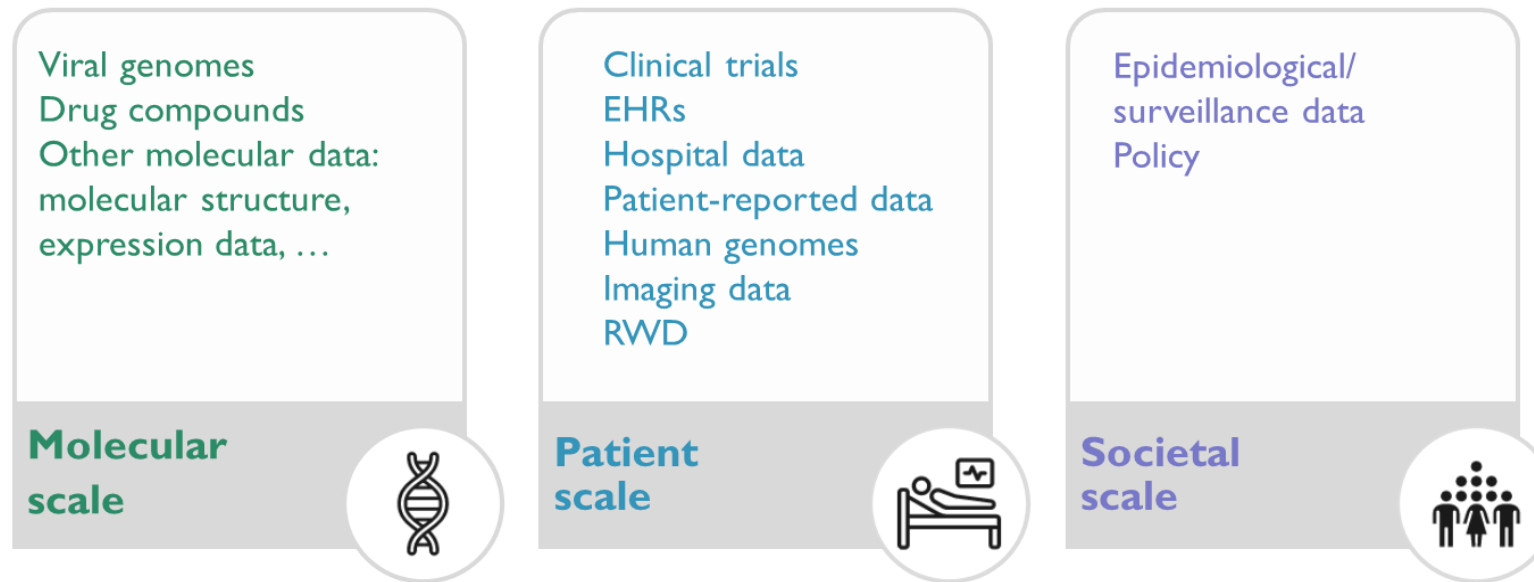| Data platforms / portals	Data science initiatives
COVID-19 collaborations	Data sharing policy

Data portals and platforms

- Molecular, patient and societal-scale data repositories, platforms, and libraries

- National portals and dashboards

Types of data

- There are **nearly 100 data repositories, platforms, databases and libraries** of datasets that are relevant to COVID-19.

- For the purpose of this analysis, we have classified data into three main types.*

Viral genomes Drug compounds Other molecular data: molecular structure, expression data, … **Molecular scale**	Clinical trials EHRs Hospital data Patient-reported data Human genomes Imaging data RWD **Patient scale**	Epidemiological/ surveillance data Policy **Societal scale**

Definitions:
- Electronic Health Records (EHRs): primary care data
- Human genomes: considered as patient rather than molecular data in consideration of privacy issues
- Real World Data (RWD): e.g., claims data, wearable data, prescription databases

*Classification adapted from https://arxiv.org/pdf/2003.11336.pdf, "Mapping the landscape of Artificial Intelligence applications against COVID-19", a paper by researchers from WHO, UN Global Pulse and the MILA-Quebec AI Institute.

Molecular scale

We found more than 20 initiatives in this space, building on robust existing infrastructures, well-established reputations and strong collaborations. Two main types of portals can be identified in this category:

Libraries of drug compounds

Focus on potential molecular compounds for therapeutic development

- COVID-19 preclinical drug development database;

- ICR canSAR integrated knowledgebase;

- E4C - EXSCALATE4COV chemical library

- COVID Open Data Portal NCATS is generating a collection of datasets by screening a panel of SARS-CoV-2-related assays against all approved drugs -> COVID-19-related drug repurposing data and experiments for all approved drugs.

- EU-OPENSCREEN ERIC COVID-19 data resources

Sequence data repositories

Focus mainly on viral sequence data.

- COVID-19 Data Portal led by EMBL-EBI and Elixir

- COVID-19 HG Initiative a bottom-up initiative set up by Institute for Molecular Medicine Finland

- 2019 Novel Coronavirus Resource (2019nCoVR) Established by the Genomics Data Center of the Chinese Academy of Sciences

- NCBI Virus Sars-CoV-2 Data Dashboard

COVID-19 *Data Portal*

The portal contains all COVID-19 datasets from EMBL-EBI: sequences, expression, proteins, structures, compounds, targets and literature. The Data Portal will be the primary entry point into the European COVID-19 Data Platform, which will also comprise of SARS-CoV-2 Data Hubs. These will be built on EMBL-EBI infrastructure and will be used by public health agencies and scientists responsible for generating viral sequences at national or regional levels.

Patient scale platforms

A number of existing data platforms for clinical trial datasets now also offer a platform for COVID-19 data and some new portals are also being established. These platforms all ensure security of data and strong governance frameworks. One of the key challenges identified by researchers is how to facilitate research across different datasets, often with difficulties of interoperability and data harmonisation.

Vivli, a non-profit organisation, established a global platform for sharing of individual participant-level trial data and associated documents in 2018. The platform offers an independent data repository, in-depth search engine and a secure research environment, within the Microsoft Azure Cloud. Vivli has now launched a COVID-19 portal for sharing of completed interventional treatment trial data.

IDDO and ISARIC have partnered to collect and share COVID-19 clinical data. ISARIC has collected standardised COVID-19 clinical data from 52 countries. IDDO has a platform to assemble clinical, laboratory and epidemiological data to be shared with the research and humanitarian communities. Researchers can request access to data in the repository and ISARIC and IDDO are developing systems to make the data more widely available.

This US-based initiative aims to build a secure repository of HIPAA-compliant, de-identified and limited patient-level data sets made available to public health and policy researchers to extract insights to help combat the COVID-19 pandemic. The database is supported by a public-private consortium, composed of institutions donating technology service, healthcare expertise, and limited and de-identified data. The cross-industry collaborative includes Datavant, Health Care Cost Institute, Medidata, Mirador Analytics, Veradigm, Change Healthcare, Snowflake and many others.

EDCTP has launched a new data repository finder, although this is not COVID specific.

Comparing data platforms

	COVID-19 Data Portal	Vivli	ISARIC IDDO	COVID-19 RESEARCH DATABASE	ICODA
Type of data	(molecular)	(clinical trial)	(patient + clinical) (not pharma)	(patient)	(patient + clinical)
Data repository	✓	✓	✓	✓	✗
Controlled access	✓	✓	✓	✓	✓
Secure analysis environment	✓	✓	✗	?	✓
Federated analysis	✓	✗	✗	✗	✓
Driver project focus	✗	✗	✗	✗	✓
Reach	Global (EU base)	Global (US base)	Global (esp. LMIC)	US	Global

Molecular level   Clinical trial data   Patient data

NOTE: This information is based on our best assessment from publicly available information.

Patient scale – other examples

WHO Global COVID-19 Clinical Data Platform	Collects anonymized clinical data relating to hospitalized cases of COVID-19, available to WHO to help inform response
Covid Alliance Research Platform	A non-profit coalition bringing together researchers and data scientists to make public-access datasets available for policy makers and the public
End Pandemic National Data Consortium	US-focused
iMove-COVID-19 EU	Repurposing the Influenza: monitoring vaccine effectiveness in Europe with H2020 funding, to pool data to answer specific research questions
HARMONY Alliance COVID-19 Data Platform	A public-private Partnership for Big Data in Hematology that is applying its expertise and network to COVID-related data collection and sharing. Data will be available through the EBI COVID-19 Data Portal.
COVID-19 Data Exchange	Launched by Dawex, and supported by a number of private partners, the Exchange offers an easy-to-use platform for public and private organizations to securely source, publish and exchange non-personal data, including statistical data, research data and test results.

National portals - examples

(includes both patient and societal scale data)

UK	COVID-19 National Core Studies Health Data Research Innovation Gateway NHS COVID-19 Data Store
Netherlands	National observational COVID-19 data portal ZonMw has commissioned Health-RI and the GO FAIR foundation to develop a data portal to help researchers find and reuse COVID-19 related observational data from Dutch health care providers
Italy	Risorse dati su COVID-19
Germany	National Research Data Infrastructure for Personal Health Data Covid-19 Taskforce
Sweden	COVID-19 Data Portal Sweden
Croatia	Croatian Open Data Portal on COVID-19
Australia	COVID-19 Clinical Data Analytics Platform, led by Queensland University of Technology, Monash University and University of Sydney
South-East Asia	Centre for Strategic and International Studies Southeast Asia COVID-19 Tracker
South Korea	DataOn platform, OpenDataforCovid-19
Sierra Leone	COVID-19 Live Update
Pan Africa	ONE Africa COVID-19 Tracker, Africa Covid-19 Dashboard
Ghana	COVID-19 Monitoring Dashboard
Colombia	COVID-19 Colombia Project Novel Coronavirus Data Repository
US	CDC COVID Data Tracker

Societal scale

- There are an increasing number of surveillance dashboards, at national, regional and global level. These are mainly based on high-level aggregate information and have been developed to track the virus in real-time and inform policy decisions.

- While many of these have user-friendly interfaces, recent commentary has highlighted that the underlying data is often disparate, messy and inconsistent, making cross-country comparisons difficult.* The real challenge, as identified by the VODAN initiative, lies with real world observations made on actual situations in real time. These data are usually collected and shared in a more fragmented and disorganised way and the fact they derive from heterogeneous sources (e.g., primary care settings, private parties) pose issues in terms of interoperability.

- A significant number of dashboards and trackers, and several collaborations that have been established to help inform decision makers through accurate information, were identified in this category.

Dashboards - examples

[Johns Hopkins Coronavirus Resource Center](#)	A continuously updated source of COVID-19 data and expert guidance. It aggregates and analyzes the best data available on COVID-19—including cases, as well as testing, contact tracing and vaccine efforts—to help the public, policymakers and healthcare professionals worldwide respond to the pandemic.
[WHO COVID-19 Dashboard](#)	Mapping the number of confirmed COVID-19 cases and deaths by country
[European Centre for Disease Control](#)	The dataset and dashboard contains the latest available public data on COVID-19 including a daily situation update, the epidemiological curve and the global geographical distribution (EU/EEA and the UK, worldwide).
[COVID-19 Mobility Data Network](#)	A network of infectious disease epidemiologists around the world working with technology companies to use aggregated mobility data to support the COVID-19 response. The goal is to provide daily updates to decision-makers at the state and local levels on how well social distancing interventions are working, using anonymized, aggregated data sets.
[COVID-19 Multi Model Comparison Collaboration](#)	Provides country governments and other model users with an overview of aims, strengths and caveats of the existing COVID-19 models and how their projections differ and what the key-assumptions and drivers are of their models.
OECD [Country policy tracker](#)	Provides information about the response to COVID-19 at a country level

Data sources

- Research trackers and trial registries

- COVID-19 datasets

Research trackers and trial registries

A number of COVID-specific research trackers and trial registries are maintaining up-to-date lists and analyses of COVID studies:

UKCDR tracker	A register of all COVID-19 research projects that have been funded across all disciplines, aiming to provide as comprehensive a picture as possible of the research landscape. The data is derived from UK and GloPID-R members
COVID-19 clinical trials registry (Cytel, also available through ICODA Workbench)	Data is pulled from the International Clinical Trials Registry Platform, including those from the Chinese Clinical Trial Registry, ClinicalTrials.gov, Clinical Research Information Service - Republic of Korea, EU Clinical Trials Register, ISRCTN, Iranian Registry of Clinical Trials, Japan Primary Registries Network, and German Clinical Trials Register.
Infectious Diseases Data Observatory (IDDO)	Living systematic review of COVID-19 clinical trial registrations and a visual tool to map them.
Cochrane	Cochrane CENTRAL Central Register of Controlled Trials Covid-19 study register
ECRIN	COVID-19 taskforce and trial registry
COVID-NMA initiative	A living mapping and living systematic review of Covid-19 trials
African trial registries	Pan African Clinical Trials Registry
South American trial registries	ReBEC Brazilian Registry of Clinical Trials REPEC Peruvian Clinical Trial Registry RPCEC Cuban Public Registry of Clinical Trials

Datasets - examples

Patient and disease registries that are collecting COVID-relevant data:

- ASH RC COVID-19 Registry for Hematology

- CAPACITY (BHF cardiovascular disease project)

- COVID-19 & Cancer Consortium

- COVID-19 Dermatology Registry

- COVID-19 Global Rheumatology Alliance

- COVID-19 in liver disease reporting registry

- MSIF Global Data Sharing Initiative

- PAN-COVID (Pregnancy and Neonatal)

- PRIORITY Pregnancy Coronavirus Outcomes Registry

- PIDTRAN USA Pediatric COVID-19 Registry

- SECURE-Cirrhosis registry

- SECURE-IBD database

- SECURE-SCD registry

- TIF Reporting of COVID-19 in haemoglobin disorders

- Cure HIV-COVID Reporting Database

- T1D Exchange: Surveillance of COVID-19 in Patients with Type 1 Diabetes

Many **cohorts and biobanks** are repurposing or collecting specific COVID-related data, e.g.:

- BBMRI-ERIC COVID-19 resources

- US National COVID Cohort Collaborative (N3C)

- UK Biobank COVID-19 Hub

- COVID-19 Quebec Biobank

Crowd-sourced information, including symptom trackers, e.g.:

- Zoe symptom tracker app

- Apple COVID-19

- Cough against COVID

- Coronagenes

- Coronaisrael

Clinical and epidemiological datasets, e.g.

- OpenSAFELY (UK)

- DECOVID UK)

- COVID-19 Research Database (US, Datavant)

- ISARIC

COVID-19 initiatives

- Research consortiums and collaborations
- Clinical academic societies
- Data science initiatives
- Data policy initiatives

COVID-19 collaborations - examples

The mapping identified more than 40 collaborations that have come together to tackle specific aspects of COVID, including genomics consortiums, clinical trial collaborations, and global health research collaborations. A significant number are public-private collaborations, but it is often difficult to determine progress with some of these consortiums.

ICODA
International COVID-19
Data Alliance

4CE Consortium for Clinical Characterization of COVID-19	An international consortium for electronic health record data-driven studies of the COVID-19 pandemic. The goal of this effort—led by the i2b2 international academics users group—is to inform doctors, epidemiologists and the public about COVID-19 patients with data acquired through the health care process.
COG-UK	COVID-19 Genomics UK Consortium
CanCOGeN	Canadian COVID Genomics Network
COVID-19 Evidence Accelerator	Brings together US experts in health data aggregation and analytics to assess RWD, working with the FDA
COVID-19 CP Collaboration Platform	US based initiative working with Vivli to encourage collaboration between RCTs, including a platform to allow sharing of RCT protocols
CAIAC	Collective and Augmented Intelligence against COVID-19
ECHAlliance	European Connected Health Alliance is the Global Connector for Digital Health with a number of pop-up ecosystems to tackle COVID-19
Trinity Challenge	Announced by Dame Sally Davies as a coalition united by the common aim of using data and advanced analytics to develop insights and actions to contribute to a world better protected from health emergencies
COVID-19 Clinical Research Coalition	Aims to accelerate COVID research in resource-limited settings
ALERRT	African Coalition for Epidemic Research, Response and Training
PANDORA	Pan-African Network for Rapid Research, Response and Preparedness for Infectious Diseases Epidemics

Clinical academic societies

Professional bodies are also actively involved in the COVID-19 response, including a focus on the importance of data sharing. Examples include:

- European Respiratory Society: END-COVID - European Respiratory Network for data sharing in COVID-19

- WHO GARD : Global Alliance Against Respiratory Diseases which has a large network and lots of resources

- Forum of International respiratory societies

Data science initiatives

A number of online platforms have been created where data scientists and ML researchers can apply their skills to address requests for help from the research and clinical community.

United against COVID	Brings together COVID OSS Help, Data against COVID and Crowdfight COVID-19
COVID OSS Help	Open source scientific computing software core developers offer their help to virologists / epidemiologists / researchers working on a cure
Data against COVID	Community connecting data scientists and bioinformaticians with medical professionals / biochemists / virologists / infectious disease specialists / drug developers
Crowdfight COVID-19	Linking volunteer researchers with COVID-19 projects that need support
COVID-19 Data Science Consortium	India-based consortium established by Wadhwani AI, focusing on modelling and diagnostic tools (see Deep Dive). Partners include Gates, the Harvard-led COVID-19 Mobility Data Network, PATH and UNICEF
COVID-19 Data Forum	A joint project of the R Consortium and the Stanford Data Science Institute, hosting a series of multidisciplinary webinars and online meetings for experts to discuss data-related aspects of the scientific response to the pandemic.
AI-Robotics vs COVID-19 Alliance	EC initiative
Data Science Africa COVID-19 Response	
DATA-POP Alliance	
Institut Français de Bioinformatique	
RECON R Epidemics Consortium	COVID-19 Challenge
COVID Alliance	

Data policy initiatives

The Global Research Collaboration for Infectious Disease Preparedness, an international network of major research funders, has highlighted the challenges of timely and transparent sharing of data during an epidemic. The data sharing working group published a Roadmap and Principles for data sharing in an epidemic situation.

Data Together: CODATA, GO FAIR, RDA and WDS have committed to work together to accelerate implementation of FAIR data ecosystem focused on COVID-19.
- Research Data Alliance published Recommendations and Guidelines on Data Sharing for COVID-19 (30/06/2020).
- CODATA has been tasked by the International Science Council to prepare a major global programme 'Making Data Work for Cross-Domain Grand Challenges' for launch in 2021, which has now expanded to COVID-19.
- The VODAN (Virus Outbreak Data Network) Implementation Network, aims to make source data FAIR and make them available for reuse in a distributed manner. The FAIRification process is initially focused on updating Clinical Research Forms (CRF) to meet the WHO standards.

Next steps

Building a living map

- With the landscape continuing to evolve rapidly, we know this list of COVID-19 data initiatives is not complete. We need your help to keep it as up-to-date and comprehensive as possible.

- If there is something missing, or you are involved in other activities that you would like to see included, please do let us know.

- This is a living document that we aim to keep up-to-date. If you would like to collaborate on updating this document, please get in touch.

International.alliance@hdruk.ac.uk