Data Sharing in Epidemiology

Version: 0.05

DOI: 10.15497/rda00049

Authors: RDA-COVID-19-Epidemiology Working Group

Co-chair: Priyanka Pillai

Moderators: Claire Austin, Gabriel Turinici

Published: 28th May 2020

Abstract: An immediate understanding of the COVID-19 disease epidemiology is crucial to slowing infections, minimizing deaths, and making informed decisions about when, and to what extent, to impose mitigation measures, and when and how to reopen society. Despite our need for evidence based policies and medical decision making, there is no international standard or coordinated system for collecting, documenting, and disseminating COVID-19 related data and metadata, making their use and reuse for timely epidemiological analysis challenging due to issues with documentation, interoperability, completeness, methodological heterogeneity, and data quality. There is a pressing need for a coordinated global system encompassing preparedness, early detection, and rapid response to newly emergent threats such as SARS-CoV-2 virus and the COVID-19 disease that it causes.

The intended audience for the epidemiology recommendations and guidelines are government and international agencies, policy and decision makers, epidemiologists and public health experts, disaster preparedness and response experts, funders, data providers, teachers, researchers, clinicians, and other potential users.

Keywords: COVID-19; Supporting output; Epidemiology; Recommendations; Data Sharing

Language: English

License: Attribution 4.0 International (CC BY 4.0)

Citation and Download: RDA COVID-19 Working Group, subgroup Epidemiology (2020): Data Sharing in Epidemiology, Version 1.0. DOI: 10.15497/rda00049
DATA SHARING IN EPIDEMIOLOGY
SUPPORTING DOCUMENT
2020-05-28

Contents
1 Focus and Description ........................................................................................................... 2

2 Scope .................................................................................................................................. 2

3 Policy Recommendations ..................................................................................................... 2
  3.1 General ............................................................................................................................. 2
  3.2 Information Technology and Data Management ............................................................... 3
  3.3 COVID-19 Epidemiological data, analysis and modeling ............................................... 4

4 Guidelines ............................................................................................................................ 5
  4.1 COVID-19 Population Level Data Sources ....................................................................... 5
  4.2 Epidemiological Surveillance Data Model ......................................................................... 5
  4.3 COVID-19 Survey Initiatives ............................................................................................. 6
  4.4 COVID-19 Question Bank ................................................................................................ 7
  4.5 Privacy .............................................................................................................................. 7
  4.6 Global Preparation, Detection and Response ...................................................................... 7
  4.7 A Common Data Model ................................................................................................... 8
  4.8 Putting It All Together: Epi-Stack ................................................................................... 8

APPENDIX 1
A Full Spectrum View of the COVID-19 data domain: An Epidemiological Data Model ................... 10
Jay Greenfield, Meg Sears, Rajini Nagrani, Claire C. Austin; and the RDA-COVID19-WG

APPENDIX 2
Epi-TRAC: Rapid detection and whole system response for emerging pathogens ................................ 13
Jay Greenfield, Henri Tonnang2, Gary Mazzaferrro, Claire C. Austin, and the RDA-COVID19-WG

APPENDIX 3
Carsten O. Schmidt, Stefan Sauermann, Rajini Nagrani, Anna Widyastuti, Chifundo Kanjala, Jay Greenfield; and the RDA-COVID19-WG

APPENDIX 4
Privacy ....................................................................................................................................... 24
Stefan Sauermann, Chifundo Kanjala, Matthias Templ

APPENDIX 5
Augmented common data models: An Epi-STACK architecture for COVID-19 epidemiology datasets .......... 28
Jay Greenfield, Rajini Nagrani, Meg Sears, Anna Widyastuti, Claire C. Austin; and the RDA-COVID19-WG

BIBLIOGRAPHY .................................................................................................................... 31-36
1 Focus and Description

An immediate understanding of the COVID-19 disease epidemiology is crucial to slowing infections, minimizing deaths, and making informed decisions about when, and to what extent, to impose mitigation measures, and when and how to reopen society.

Despite our need for evidence based policies and medical decision making, there is no international standard or coordinated system for collecting, documenting, and disseminating COVID-19 related data and metadata, making their use and reuse for timely epidemiological analysis challenging due to issues with documentation, interoperability, completeness, methodological heterogeneity, and data quality.

2 Scope

There is a pressing need for a coordinated global system encompassing preparedness, early detection, and rapid response to newly emergent threats such as SARS-CoV-2 virus and the COVID-19 disease that it causes.

The intended audience for the epidemiology recommendations and guidelines are government and international agencies, policy and decision makers, epidemiologists and public health experts, disaster preparedness and response experts, funders, data providers, teachers, researchers, clinicians, and other potential users.

3 Policy Recommendations

3.1 General

1. Urgently update data sharing policies and Memoranda of Understanding (MOUs) across all domains, in government, healthcare systems, and research institutions to support Open Data, Open Science, scientific data modernization, and linked data life cycles that will enable rapid and credible scientific and epidemiologic discovery, and fast-track decision-making.

2. Streamline data flow between sub-national jurisdictions/institutions and their national government, and countries and international organizations.

3. Implement a “data first” publication policy in research by treating publication of data articles in “open” peer-reviewed data journals, including the deposit of data and associated code in a trusted digital repository with tiered access to appropriately credentialed people and machines to preserve data security.
4. Peer-reviewed data articles should be treated as first-class research outputs equal in value to traditional peer-reviewed articles.

5. Call upon the international Open Government Partnership (OGP) to add “Open Science” as one of its Policy Areas to be included in National Action Plans. Hold member countries accountable for developing and implementing Open Science commitments.

6. Publish situational data, analytical models, scientific findings, and reports used in decision-making and justification of decisions (OGP 2020).

3.2 Information Technology and Data Management

1. Invest in state-of-the-art information technology (IT) and data management system infrastructure (devices, hardware, algorithms, software used to store, retrieve and process data).
   a. Rapid development of a modern data management system infrastructure will ensure scientific data integrity via data management plans embedded in linked data life cycles that: (a) are fully machine-enabled, and not constrained by non-digital processes; (b) are available online end-to-end; (c) enable synchronous and asynchronous workflows; (d) guarantee tidy, Findable, Accessible, Interoperable, Reusable, Ethical, and Reproducible (FAIRER) data, metadata, and code/scripts; (e) guarantee data security; (f) provide tiered access to restricted data by appropriately credentialed users and machines; and, (g) analytical tools.
   
   b. When evaluating apps, consider the many underlying issues: legal, confidentiality, data completeness, representativeness, data quality, reliability, verifiability, data ownership, data access, data openness, data control, transparency, peer-review, etc.
   
   c. In resource limited settings, leverage the use of existing data management system infrastructure while still ensuring data quality and integrity.
   
   d. explore strategies that will facilitate rapid data sharing within and between government and international agencies, other organizations, institutions and individuals (i.e. specific data users).

2. Document all methodologies used to collect, define, compile, and analyze data, including data management, data cleaning, data quality checks, updating, data imputation, computer code used, definitions used, etc.

3. Ensure an appropriate semantic annotation of data to facilitate its comparability across studies and countries, using as much as possible established standards (e.g. LOINC, UMLS).
4. Rapidly develop standardized tools for aggregating microdata to a harmonized format(s) that can be shared and used while minimizing the re-identification risk for individual records.

5. Develop machine readable citations and micro-citations for dynamic data. Rapid development of: (a) Resolvable Persistent Identifiers, rather than Uniform Resource Locators (URLs); (b) Machine readable citations; (c) Micro-citations that refer to the specific data used from large datasets; and, (d) Date and Time Access citations for dynamic data (ESIP 2019).

3.3 COVID-19 Epidemiological data, analysis and modeling

1. Rapidly develop a consensus standard for COVID-19 surveillance data:
   a. Definition of and reporting criteria for COVID-19 testing, reporting on testing, and testing turnaround times.
   
      b. Policies and definitions: interventions, contact tracing, reporting of cases, deaths, hospitalizations and length of stay, ICU admissions, recoveries, reinfections, time from contact if known, symptoms onset and detection, through clinical course and interventions, to death or recovery, comorbidities, long-term effects in recovered cases, sequelae and immunity, location, demographic, socioeconomic information, and outcome of resolved cases.

   c. Uniform standard daily reporting cut-off time.

2. Rapidly develop an internationally harmonized specification to enable the export/import/integrate epidemiologic data across different levels of data generation (e.g., clinical systems, population-based surveillance/research data, data from biomarker and omics studies, death certification, health insurance data), and successful record-linkage.

3. Develop systems that support workflows to link and share data between different domains, while protecting privacy and security. Use domain specific, time stamped, encrypted person identifiers for this purpose.

4. Implement internationally harmonized COVID-19 intervention protocols based on peer-reviewed empirical modeling and epidemiological evidence, considering local conditions.

5. Account for public health decision making demands in COVID-19 studies.

6. Harmonize approaches to comparably assess and quantify side-effects of pandemic containment and mitigation measures.
7. Report underlying assumptions and quantify effects of uncertainties on all reported parameters and conclusions for all model predictions etc.

8. Implement a data-driven approach for early identification of hotspots.

4 Guidelines

4.1 COVID-19 Population Level Data Sources

Although jurisdictions within countries send COVID-19 population level data to the national level, and member countries send data to the WHO, other organizations also collect COVID-19 surveillance data from various sources for a variety of reasons (Table 1). Epidemiologists are thus faced with a situation where it is difficult to assess which datasets are the most up to date, complete and reliable.

<table>
<thead>
<tr>
<th>Table 1. COVID-19 population level data sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Centre for Disease Control Geographic distribution of COVID-19 cases worldwide</td>
</tr>
<tr>
<td>European Centre for Disease Control The European Surveillance System (TESSy)</td>
</tr>
<tr>
<td>Institute for Health Metrics and Evaluation (IHME) Global Health Data Exchange (GHDx)</td>
</tr>
<tr>
<td>Johns Hopkins University COVID19 dataset</td>
</tr>
<tr>
<td>Oxford University COVID19 dataset</td>
</tr>
<tr>
<td>The Atlantic COVID Tracking Project</td>
</tr>
<tr>
<td>The New York Times Covid-19 Data in the United States</td>
</tr>
<tr>
<td>The White House COVID-19 Open Research Dataset Challenge (CORD-19)</td>
</tr>
<tr>
<td>U.S. Centre for Disease Control Cases of COVID19 in the U.S.</td>
</tr>
<tr>
<td>University of Washington Be Outbreak Prepared</td>
</tr>
<tr>
<td>World Bank Understanding the Coronavirus (COVID-19) pandemic through data</td>
</tr>
<tr>
<td>World Health Organization (WHO) Novel Coronavirus (2019-nCoV) situation reports</td>
</tr>
<tr>
<td>Worldometer COVID19 data</td>
</tr>
</tbody>
</table>

4.2 Epidemiological Surveillance Data Model

The COVID-19 epidemiology that guides public health decisions is dependent on interoperable input data from across a wide variety of domains that include not only clinical, surveillance, research, and modelling data, but also administrative, demographic, socioeconomic, cultural practices and lifestyle, and environmental data, amongst others.

An epidemiological surveillance data model must include the primary data domains that need to be integrated to understand COVID-19, and to improve surveillance and follow-up: (a) clinical event history and disease milestones; (b) epidemiological indicators and reporting data; (c) contact tracing; (d) personal risk factors.
Standardization challenges within each of these domains remain to be solved before data can be effectively integrated across domains for epidemiology studies. For example, on the clinical side, the U.S. Clinical Data Interchange Standards Consortium (CDISC) new specification (Interim User Guide for COVID-19), and the WHO Core and Rapid COVID-19 Case Reporting Forms used in low- and middle-income Countries (LMIC) require additional harmonization.

4.3 COVID-19 Survey Initiatives

International efforts are currently underway to create COVID-19 instruments/questionnaires (Tables 2 and 3). These COVID-specific tools are concentrated at person-level for clinic/hospital surveillance (e.g., Case Report Forms-CRFs), or community surveillance (e.g., questionnaire for general population), and do not necessarily collect the same data. Adherence of new studies to already introduced instruments will strongly enhance the comparability of results.

**Table 2. Questionnaire instruments: Reference studies**

<table>
<thead>
<tr>
<th>CLINICAL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia:</td>
<td>NSW Case questionnaire</td>
</tr>
<tr>
<td>Germany:</td>
<td>Covid-19 research dataset</td>
</tr>
<tr>
<td>Uganda:</td>
<td>Perinatal COVID-19 Uganda</td>
</tr>
<tr>
<td>US:</td>
<td>Human Infection with 2019 Novel Coronavirus Person Under Investigation (PUI) and Case Report Form</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WORLDWIDE</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>POPULATION-BASED</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Brazil:</td>
<td>Brazil Prevalence of Infection Survey</td>
</tr>
<tr>
<td>Europe:</td>
<td>Questionnaire by WHO Europe</td>
</tr>
<tr>
<td>Germany:</td>
<td>GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany</td>
</tr>
<tr>
<td>Germany:</td>
<td>NAKO COVID-19 Survey tool</td>
</tr>
<tr>
<td>Israel</td>
<td>One-minute population wide survey</td>
</tr>
<tr>
<td>Low and Middle Income Countries:</td>
<td>LMIC Covid Questionnaire</td>
</tr>
<tr>
<td>South Africa:</td>
<td>South African Population Research Infrastructure (SAPRIN) COVID-19 Screening Form</td>
</tr>
<tr>
<td>South Asian countries:</td>
<td>National Institute for Health Research (NIHR) Global Health Research Unit</td>
</tr>
<tr>
<td>UK</td>
<td>UK COVID-19 Questionnaire</td>
</tr>
<tr>
<td>Worldwide (WHO):</td>
<td>Population-based age-stratified sero-epidemiological investigation protocol for COVID-19 virus infection</td>
</tr>
</tbody>
</table>
4.4 COVID-19 Question Bank

Some of the questionnaire initiatives shown in Tables 2 and 3 are currently feeding into the construction of a COVID-19 demographic and epidemiological surveillance question bank that can be used to form locality specific surveys with both common and distinct questions by domains and cohorts (Wellcome Trust). Some, such as the UK COVID-19 Questionnaire or the Covid-19 research dataset are now being funded. Question banks, once they become operational, can be queried and filtered by domain, cohort, question text, etc. Based on such queries, new questionnaire products can be developed that are more or less interoperable, depending on the questions selected and the capture of “localization” information in the question metadata when questions are reused from one survey to the next.

4.5 Privacy

Data sharing is essential to improve epidemiological analyses, cross-border pandemic modeling, and coordinated policy development between countries. To ensure privacy, both pseudo-anonymization of direct identifiers (e.g. patient specific ID’s) and anonymization of indirect identifiers (e.g. socio-demographic information on individuals) must be applied. In addition, it is necessary to control statistical disclosure risk to prevent identification of individuals and their health status using a combination of indirect identifiers such as education level, sex, age, clinical conditions, among others (Duncan et al. 2011; Templ et al. 2015; Templ 2017). Using synthetic data may be an option to lower re-identification risks while retaining properties of the original data sets.

4.6 Global Preparation, Detection and Response

WHO’s Global Influenza Surveillance Response System (GISRS) is a well-established network of more than 150 national public health laboratories in 125 countries that monitors the epidemiology and virologic evolution of influenza disease and viruses (WHO 2020). Prior to the COVID-19 outbreak, WHO was already engaged in re-examining GISRS’s long-term fitness-for-purpose. In line with these short-term considerations and with GISRS long-term aspirations, we are recommending a real time, adaptable, rapid response system that supports developing countries, and that employs new technology to combat pandemics and other emerging diseases. The RDA-COVID19-Epidemiology WG recommends the creation of a WHO-led EPIdemiological Translational Research Action Coalition (Epi-TRAC) to add an implementation
layer to the existing WHO policies, guidelines, partnerships, and information exchange stack adapted to country-specific contexts.

4.7 A Common Data Model

Data models may make use of the broad ecosystem of surveillance and clinical data that can also include contact tracing apps, biospecimens, and environmental sample data collected in the community/population or clinic/hospitals.

An emulated trials approach may enable assessment of various risk and prognostic factors (Hernan et al.). Application of a Common Data Model (CDM) for COVID-19 would facilitate comparing clinical burden and patient outcomes in the context of previous environmental and exposures and comorbidities.

Another possible use case is decision support following an early warning system alert of emergence of a novel pathogen such as SARS-CoV-2. The CDM provides a framework for making public health policy decisions, using partial information about the pandemic that leverages population-level population and health information, person-level epidemiological surveillance information collected in the field and, at the same time or alternatively, person-level patient care information collected in a clinic or hospital setting.

4.8 Putting It All Together: Epi-Stack

The WHO has established the Information Network for Epidemics (Epi-WIN) covering four strategic areas: (a) Identify; (b) Simplify; (c) Amplify; and, (d) Quantify. Evidence is gathered, appraised, and assessed to help form recommendations and policies that have an impact on the health of individuals and population.

The RDA Epi subWG proposes an expanded Epi-Stack feeding into Epi-WIN (Figure 1). This would bring together in a managed system a common data model, the epidemiological surveillance data model, clinical and questionnaire data, population level indicators, and core use cases (Epi-TRAC early warning and response system, decision support research, and patient care research).
Figure 1. Epi-Stack. Proposed evidence support system as input to the WHO’s Epi-WIN communication channels for various audiences.
APPENDIX 1 - A Full Spectrum View of the COVID-19 data domain: An Epidemiological Data Model

Jay Greenfield1,*, Meg Sears2, Rajini Nagrani3, Claire C. Austin4; and the RDA-COVID19-WG5

1Data Documentation Initiative, 2Ottawa Hospital Research Institute, 3Leibniz Institute for Prevention Research and Epidemiology-BIPS, 4Environment and Climate Change Canada, 5This work is linked to the Research Data Alliance RDA-COVID19-WG. Epidemiology subWG recommendations and guidelines on data sharing.

*Corresponding author
All views and opinions expressed are those of the co-authors, and do not necessarily reflect the official policy or position of their respective employers, or of any government, agency or organization.

Background
In a full spectrum COVID-19 domain model, it is necessary to account for the entire individual experience from susceptibility to exposure to infection and, through treatment, to death or recovery with and without sequela [Pesquita C, Ferreira JD, Couto FM, Silva MJ, 2014] [World Health Organization, 2020 Apr 23] [World Health Organization, 2020 Mar 24]. Susceptibility includes person-level risk factors, the toolbox of public health countermeasures in play and the use a person is able to make of these countermeasures. In the full spectrum model susceptibility interfaces with exposure through contacts that may be captured prospectively and/or retrospectively. Diagnosis may follow or not and can take many forms. Treatment may or may not follow diagnosis. Treatment may or may not include hospitalization. Death may occur without hospitalization. And recovery may be fraught with sequela. The full spectrum domain model presented here accounts for all of these pathways.

Methods
The full spectrum model is based on the SEIR (Susceptibility, Exposure, Infection, Recovery) domain model [Sun, P and Kang L, 2020]. We augmented the SEIR model to take into account public health countermeasures [Giordano G, Blanchini F, Bruno R et al, 2020] and differences in the availability of resources between High Income Countries (HIC) and Low and Middle Income Countries (LMIC). The augmented SEIR domain model might be more informed by the hospital experience in HICs, and by community-based demographic and epidemiological surveillance in LMICs [Wang C, Pan R, Wan X, et al, 2020]. Some LMICs also benefit more directly from lessons learned from experiences with Ebola and HIV AIDS. The augmented SEIR model gives equal weight to both the hospital experience and to the community-based demographic and epidemiological surveillance.

1 In ontology engineering, a domain model is a formal representation of a knowledge domain with concepts, roles, datatypes, individuals, and rules, typically grounded in a description logic.
Results

Figure 1. Epidemiology surveillance data model. This is a full spectrum domain model for COVID-19.
Discussion
The full spectrum view is longitudinal in scope. An essential use of this model is as the basis for a specification for an implementation data model that supports a full spectrum view of an individual’s COVID-19 experience. It drills down to a granular level with the hospitalization experience, and with an individual’s COVID-19 experience in the community before and after hospitalization, or in the absence of hospitalization. Because the full spectrum view takes into account event histories, it includes repeat visits to clinics and hospitals, changed diagnosis, and repeat contacts with field workers.

Author roles
All authors accept responsibility for the content of the article
The same author may have multiple roles, and multiple authors may share a single role.
Conceptualization: JG, MS, CCA
Methodology: JG
Investigation: JG, RN, MS
Validation: RN
Writing: JG
Visualization: JG

References
World Health Organization. COVID-19 Core Version CRF. Core COVID-19 CRF - English. 2020 April 23
APPENDIX 2 – Epi-TRAC: Rapid detection and whole system response for emerging pathogens

Jay Greenfield1,*, Henri Tonnang2, Gary Mazzaferro3, Claire C. Austin4; and the RDA-COVID19-WG*

1Data Documentation Initiative, 2International Center of Insect Physiology and Ecology (ICIPE), Kenya, 3Independent researcher, 4Environment and Climate Change Canada, 4This work is linked to the Research Data Alliance RDA-COVID19-WG, Epidemiology subWG recommendations and guidelines on data sharing.

*Corresponding author

All views and opinions expressed are those of the co-authors, and do not necessarily reflect the official policy or position of their respective employers, or of any government, agency or organization.

Background

COVID-19 threat detection has been slow relative to the speed with which the epidemics and subsequent pandemic spread. As a result, countries have implemented a series of severe public health measures that have differed from country to country depending on a variety of factors.

Although high income countries (HIC) have some form of early response system, the tempo, stealth, and spread of this novel coronavirus left most countries with little time to act leading to widespread lockdowns to suppress or mitigate the spread [Gates B, 2020; Bai Z, et al. 2020].

Lockdowns in turn have economic consequences for both HICs and LMICs. We are learning how to phase and stage the public health and economic health toolboxes within and between regions and countries as part of a single response system nationally and internationally.

Methods

In light of the COVID-19 pandemic experience, both the early warning and response parts of our systems need to be rethought. This is necessarily occurring in the midst of the pandemic. In fact, varying responses around the world have created a number of natural experiments: South Korea, Sweden, Kenya, Brazil, California [Zwald ML, et al. 2020], South Dakota, and New York State [Cummings MJ, Baldwin MR, Abrams D, et al. 2020]. Each of these jurisdictions have implemented emergency public health and economic measures in unique ways.

Data collected from each of these natural experiments will help us to identify sentinels ("canaries in the coal mine") and response measures that prove useful or not under various conditions.
Results
We propose an Epidemiology Translational Research Action Coalition (Epi-TRAC) led by the WHO (Figure 1).

“Holistic” response thinking, where both emergency public health and economic measures are represented in a single “causal loop,” may be emerging as a methodology [Bradley et al. 2020]. In Figure 2, we show interactions that may occur during the COVID-19 pandemic between the economic sector, the healthcare system, and delays.
Figure 2. A notional COVID-19 emergency public health and economic measures causal loop.

Discussion
“Inclusive Growth and Recovery Management” is a challenge, with a call for entries being supported by the Rockefeller Foundation [The Rockefeller Foundation, 2020]. Here and elsewhere “data science breakthroughs” are being counted upon to inform the development of early warning and response systems for COVID-19, future novel viruses and other potential threats characterized by breathtaking tempos.

Author roles
All authors accept responsibility for the content of the article
The same author may have multiple roles, and multiple authors may share a single role.
Conceptualization: GM, JG; Methodology: JG, HT; Investigation: CA; Formal analysis: GM, HT; Writing: JG; Bibliographic review and analysis: JG; Visualization: JG, HT
References


The Rockefeller Foundation. Call for entries: Data Science Breakthroughs for an Inclusive Recovery. [https://www.rockefellerfoundation.org/blog/call-for-entries-data-science-breakthroughs-for-an-inclusive-recovery/?utm_source=Rockefeller+Foundation+eAlerts&utm_campaign=3e2ba36552-EMAIL_CAMPAIGN_2020_MAY_TSG_LAUNCH&utm_medium=email&utm_term=0_6138ee88b7-3e2ba36552-215466361&goal=0_6138ee88b7-3e2ba36552-215466361&mc_cid=3e2ba36552&mc_eid=20f10e51cc, 2020](https://www.rockefellerfoundation.org/blog/call-for-entries-data-science-breakthroughs-for-an-inclusive-recovery/?utm_source=Rockefeller+Foundation+eAlerts&utm_campaign=3e2ba36552-EMAIL_CAMPAIGN_2020_MAY_TSG_LAUNCH&utm_medium=email&utm_term=0_6138ee88b7-3e2ba36552-215466361&goal=0_6138ee88b7-3e2ba36552-215466361&mc_cid=3e2ba36552&mc_eid=20f10e51cc)

Carsten O. Schmidt1,*, Stefan Sauermann2, Rajini Nagrani3, Anna Widyastuti4, Chifundo Kanjala5, Jay Greenfield6; and the RDA-COVID19-WG7

1University Medicine Greifswald, 2UAS Technikum Wien, 3Leibniz Institute for Prevention Research and Epidemiology-BIPS, 4Hasselt University, 5London School of Hygiene and Tropical Medicine, 6Data Documentation Initiative, 7This work is linked to the Research Data Alliance RDA-COVID19-WG, Epidemiology subWG recommendations and guidelines on data sharing.

*Corresponding author

All views and opinions expressed are those of the co-authors, and do not necessarily reflect the official policy or position of their respective employers, or of any government, agency or organization.

Background

The COVID-19 pandemic has raised a huge demand for patient and population-based data which cover not only the onset and course of the disease, and related treatments, but also side effects due to pandemic mitigation measures. For this purpose, all around the world, new instruments are being developed in a very short time. Heterogeneity between these instruments may limit comparability of results across studies and countries. Unawareness of already existing options for surveying COVID-19 related factors contributes to this heterogeneity. The present study provides an overview of some preselected instruments from different countries to guide the creation of novel tools by reusing existing instruments and items in relation to COVID-19.

Methods

We searched for publicly available online resources from major organizations in different countries to identify instruments or item banks of potential interest. Inclusion criteria for an instrument or item bank were: (1) directly COVID-19 related by targeting health-related behaviours, symptoms, treatments, clinical outcomes, pandemic mitigation measures, or COVID-19 related attitudes; (2) issued by an initiative or authority of enough weight to impact COVID-19 research, at least at the national level; (3) the instruments or item appears to be methodologically sound; (4) major parts of the instrument are of sufficient general applicability to be used beyond a local context; (5) the instrument is applicable for patient, or for general population surveys; and, (6) the instrument may be used for future research without any necessity for negotiating a license agreement. The main focus was on instruments issued in English. However, other major languages were also considered.

We described the selected instruments, and compared them at the module-level based on the targeted topic. We also referenced resources that collect instruments.
Results

Table 1a provides an overview of the instruments selected for review. Two broad categories were distinguished: (1) symptom and treatment oriented instruments for patient populations in clinic/hospital surveillance; and, (2) instruments addressing the psychosocial impact of COVID-19 and pandemic mitigation measures, and health related behaviour and attitudes in population-based surveys. Many of the latter instruments also make brief reference to symptoms or treatments. Most instruments were available only as a pdf file, while only one resource conducted an in depth semantic annotation of all items (1). There was a lack of machine readable formats to facilitate the incorporation of items and instruments into new data dictionaries. Three instrument resource pages were identified that provide access to a wider range of instruments (Table 1b).

Table 1. Questionnaire instruments: Reference studies

<table>
<thead>
<tr>
<th>Country</th>
<th>Initiative</th>
<th>Target population</th>
<th>Development stage</th>
<th>Language</th>
<th>Provenance (Influenced by...)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Australia</td>
<td>NSW Case questionnaires</td>
<td>Patients</td>
<td></td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Brazil</td>
<td>Brazil Prevalence of Infection Survey</td>
<td>Rapid tested, Tested positive</td>
<td>In development</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Europe</td>
<td>Questionnaire by WHO Europe</td>
<td>General population</td>
<td></td>
<td>German, Russian</td>
<td>Single Instrument</td>
<td></td>
</tr>
<tr>
<td>France</td>
<td>Barometer Covid19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The &quot;barometer Covid19&quot; is a citizen science initiative driven by the Datacovid association. It aims to provide open-access data from a weekly survey to</td>
</tr>
<tr>
<td>Country</td>
<td>Initiative</td>
<td>Target population</td>
<td>Development stage</td>
<td>Language</td>
<td>Provenance (Influenced by...)</td>
<td>Comments</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
<td>--------------------</td>
<td>-------------------</td>
<td>----------</td>
<td>------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>France</td>
<td>French COVID-19</td>
<td>Patients</td>
<td>In use</td>
<td>English</td>
<td></td>
<td>illuminate the struggle against the epidemic Covid19 from observations on its dynamics, its determinants and its impacts.</td>
</tr>
<tr>
<td>Germany</td>
<td>Covid-19 research dataset</td>
<td>Patients</td>
<td>In development</td>
<td>German</td>
<td></td>
<td>Lead by REACTing consortium in collaboration with ISARIC consortium (International Severe Acute Respiratory and emerging Infection Consortium)</td>
</tr>
<tr>
<td>6 Germany</td>
<td>GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany</td>
<td>General population</td>
<td>In use</td>
<td>German</td>
<td></td>
<td>National Network of German University Clinics to study COVID19.</td>
</tr>
<tr>
<td>6 Germany</td>
<td>NAKO COVID-19 Survey tool</td>
<td>General population</td>
<td>In use</td>
<td>German</td>
<td></td>
<td>As the largest European infrastructure institute for the social sciences GESIS provides essential and internationally relevant research-based services</td>
</tr>
</tbody>
</table>

The German National Cohort (GNC) has been inviting adults aged between 20 and 69 to 18 study centers throughout Germany since 2014 with more than 200,000 participants. The COVID-19 questionnaire
<table>
<thead>
<tr>
<th>Country</th>
<th>Initiative</th>
<th>Target population</th>
<th>Development stage</th>
<th>Language</th>
<th>Provenance (Influenced by...)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>may be requested at the NAKO study office (Bohn/Panreck).</td>
</tr>
<tr>
<td>7 Israel</td>
<td>One-minute population wide survey (Israel)</td>
<td>Israeli population</td>
<td>In use</td>
<td>Hebrew, Arabic, Russian, Spanish, French, English</td>
<td></td>
<td>Participants asked to fill it out on a daily basis and separately for each family member, including members who are unable to fill it out independently (e.g., children and older people).</td>
</tr>
<tr>
<td>8 Low and Middle Income Countries</td>
<td>LMIC Covid Questionnaire</td>
<td>In development</td>
<td></td>
<td></td>
<td>UK COVID-19 questionnaire, SAPRIN COVID-19 screening form</td>
<td></td>
</tr>
<tr>
<td>9 South Africa</td>
<td>South African Population Research Infrastructure (SAPRIN COVID-19)</td>
<td>In development</td>
<td></td>
<td>English, Afrikaans</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20
<table>
<thead>
<tr>
<th>Country</th>
<th>Initiative</th>
<th>Target population</th>
<th>Development stage</th>
<th>Language</th>
<th>Provenance (Influenced by...)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 Uganda</td>
<td>Screening Form</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 UK</td>
<td>Perinatal COVID-19 Uganda</td>
<td>Women pre-/perinatal</td>
<td></td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 UK</td>
<td>UK COVID-19 Questionnaire</td>
<td>Adult Resident, Children, Key worker, Partner.</td>
<td>In development</td>
<td>English</td>
<td>NIHR Global Health Research Unit</td>
<td>- World Bank code book for metadata. - Becoming a model for some African countries</td>
</tr>
<tr>
<td>12 UK</td>
<td>National Institute for Health Research (NIHR) Global Health Research Unit</td>
<td>Telephone sample</td>
<td></td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 US</td>
<td>Human Infection with 2019 Novel Coronavirus Person Under</td>
<td>Patients</td>
<td>In use</td>
<td>English</td>
<td>CDC</td>
<td></td>
</tr>
<tr>
<td>Country</td>
<td>Initiative</td>
<td>Target population</td>
<td>Development stage</td>
<td>Language</td>
<td>Provenance (Influenced by...)</td>
<td>Comments</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>------------------</td>
<td>----------</td>
<td>-------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>US/WHO</td>
<td>Population-based age-stratified seroepidemiological investigation protocol for COVID-19 virus infection</td>
<td>Patients</td>
<td>In use</td>
<td>English</td>
<td>WHO</td>
<td>This protocol has been designed to investigate the extent of infection, as determined by seropositivity in the general population, in any country in which COVID-19 virus infection has been reported.</td>
</tr>
</tbody>
</table>

Footnotes (Domains)

1. Clinical symptoms, Disease outcome, Exposure sites, Pre-existing conditions, Risk history, Sociodemographics.
2. Demographics, Home life, Test results, Transportation
3. Affect, Behaviour, Conspiracies (perceptions), COVID-19 risk perception: probability and severity, Fairness (perceptions), Frequency of Information, Influenza risk perception: probability and severity, interventions (perceptions), Knowledge and self-assessed adherence to prevention measures, Knowledge incubation, Knowledge symptoms/treatment, Lifting restrictions (pandemic transition phase), Policies, Preparedness and perceived self-efficacy, Prevention – own behaviours, Resilience (perceptions), Risk group, Rumors (open-ended), Self-assessed knowledge, Socio-demography, Trust in institutions (perceptions), Trust in sources of information, Use of sources of information, Worry.
4. Clinical symptoms, Complications, Imaging, Laboratory markers, Medical treatments, Medication, Sociodemographics.
5. Adherence to risk minimization measures, Changes in lifestyle factors, COVID infections and testing, COVID tracing, General health status, Physical symptoms, Respiratory infections, Workplace/changed employment situation.
6. Changed employment situation, Childcare obligations, Risk perception, Evaluation of political measures & their compliance, Media consumption, Risk minimization measures, Trust in politics and institutions.
6b. NAKO: General health status, respiratory diseases, testing for COVID, social contacts, adherence to pandemic mitigation measures, workplace, economic and social impact, depression, psychosocial complaints, physical activity, alcohol drinking, smoking habits.
7. Age, Geographic location (city and street), Isolation status, Sex, Smoking habits.
8. Actions in response to COVID-19, Bounded structure, Eligibility for testing, Epidemiological risk, Household enumeration, Household impact, Quarantine and hygiene directions, Symptom screen, Travel and movement (mobility), Travel history, Visit attempts.
10. Disease outcome, Sociodemographics, Symptoms.
11. Accommodation type, Away from home environment, Behavior changes, Change in benefits, Digital access, Economic activity before and after lockdown, Environmental attitudes, Environmental impact, Family relations, Financial impact, Food security, Impact on employment, Knowledge, Medication, Mental health, Mental health, Physical health, Pre-existing conditions, Social impact, Symptoms, Volunteering.
12. COVID-19 Interventions, Current living conditions, Displacement and mobility, Economic impacts, Impact of COVID-19 on health-related behaviors, Mental health, Precautions, Pre-existing conditions, Social aid, Social impact, Symptoms, Treatments for pre-existing conditions.
13. Diagnostic testing procedures, Clinical course, Medical history, Pre-existing conditions, Risk exposure, Sociodemographics, Symptoms, Treatments.
Table 2. Questionnaire instruments - Resources

<table>
<thead>
<tr>
<th>Provider</th>
<th>Initiative</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td>NIH Public Health Emergency and Disaster Research Response (DR2)</td>
<td>Diverse</td>
</tr>
<tr>
<td>NIH</td>
<td>COVID-19OBSSR Research Tools</td>
<td>Diverse</td>
</tr>
<tr>
<td>PhenX is funded by the National Institutes of Health (NIH) Genomic Resource Grant</td>
<td>PhenX COVID-19 Toolkit</td>
<td>Diverse</td>
</tr>
</tbody>
</table>

Instruments used in a clinical context

Table 2 Covid19 CRFs Content Overview (work draft Gsheet)

Discussion

A large number of COVID-19 related instruments have been developed in very a period of short time. We did not attempt to conduct a complete survey. Rather, selected instruments, item banks and resources of estimated high value in the field were identified and compared. This comparison revealed a very large scope of areas targeted by these instruments in relation to COVID-19, ranging from proximal measures such as symptoms of the disease to distal measures such as political attitudes. Amongst population-based instruments, some questionnaires were an add-on module to already existing cohort studies, while others were standalone surveillance questionnaires. Researchers must take into account the scope and feasibility of their surveys before implementing the questionnaires in specific situations/communities.

Although there have been some efforts taken to make clinical data interoperable, population-based data still require considerable attention. Finally, if applicable in a specific research context, our recommendation is to adhere as much as possible to existing items and instruments to facilitate the later comparison of results. New instruments should be made publicly available and published in a findable, reusable, and machine-readable format with appropriate semantic annotation.
APPENDIX 4 – Privacy
Stefan Sauermann, Chifundo Kanjala; Matthias Templ; and the RDA-COVID19-WG

1UAS Technikum Wien, 2London School of Hygiene and Tropical Medicine, 3Zurich University of Applied Sciences, 4This work is linked to the Research Data Alliance RDA-COVID19-WG, Epidemiology subWG recommendations and guidelines on data sharing

Anonymization and Pseudonymization
To ensure privacy, both pseudo-anonymization of direct identifiers (e.g. patient specific ID’s) and anonymization of indirect identifiers (e.g. socio-demographic information on individuals) must be applied. While pseudo-anonymization is trivial by using salted and hashed values (using e.g. the sha3 hash function and 256 bit salts) and different domains-specific salts for the same patient and a central agency that stores the salts and is able to link different data sets from with the same patients. An appropriate IT-infrastructure is relevant here for ID management and linking data.

Patient related data are recorded and used in different domains, for example medicine, communities, research, administration, and statistics. Patient specific electronic identifiers (IDs) link specific information to the individual patient records in each domain. Each domain assigns a domain-specific ID to each individual patient. In order to satisfy privacy requirements on direct identifiers, data that carry a domain ID remain within the home domain and are not shared. If data are re-used, for example in research or public health purposes, the domain-specific ID is removed (anonymization) or replaced by a different ID (pseudonym). Pseudonyms can later be traced back to the original domain ID. This must only occur under well-defined conditions, e.g. if a statistics department needs to clear ambiguities in incoming information together with the organisation that generated the data. In multi-domain and multi-organisation scenarios, consistent management of IDs and pseudonyms is needed to enable cross-linking of data from different sources using the ID’s.

The following requirements apply:
- Patient IDs exist for each domain;
- The local domain ID must not leave the local domain in clear text, to prevent unintended record linkage between domains;
- When providing data from a source domain to a target domain, the target domain patient ID (pseudonym) must become available to users in the target domain; and,
- Domain IDs must enable domains to cooperate e.g. for clearing ambiguities, while preserving privacy.

In Austria, for example, eGovernment legislation and IT infrastructures are in place to handle domain specific identifiers e.g. for health care, traffic, taxes, and statistics (reference). This is implemented and in operation for example in the Austrian electronic health care record ELGA.
Figure 1A describes how data from a health domain can be linked to records in a research domain in this way. Figure 1B introduces the needed IT infrastructure. Figure 1C shows how IDs are mapped between domains while preserving pseudonymization.

Figure 1A. Sharing or linking a body temperature observation from the healthcare domain with a research and administration domain. In the healthcare domain ID (bPK_Health, green), a patient is identified with a specific ID, (1234, colour green, denoting that it is unencrypted). A doctor will receive this data together with the original ID, as the law allows doctors to share IDs unencrypted. The doctor can attach an encrypted ID (#1hsId8, red, denoting it is encrypted) to the data. A researcher who receives the data, decrypts the encrypted ID. This decrypted ID (6742, blue, denoting that it is a pseudonym) is specific to the research domain (bPK_research, blue). The same method applies as data are provided to administrations, e.g. for public health purposes.
Figure 1B. IT infrastructure for cross-domain IDs management. A user in the medical domain queries the IT service using the ID of the health domain, asking for encrypted IDs of other domains. The service responds with the encrypted IDs. Users in other domains can use the same mechanism. This enables users in different domains to co-operate: For example, the researcher can attach the encrypted bPK_Health ID to a message to the doctor, asking for details to support clearing ambiguities in the data the doctor provided earlier. The doctor can then decrypt the patient ID, access the patient related information in the health IT system, and finalize the clearing with the researcher.

Controlling statistical disclosure risk
A pseudo-anonymized data set does not imply that individuals cannot be re-identified and, additionally, the data set still needs to be anonymized using different methods than those regarding pseudo-anonymization. Intruders can reveal patients’ identities using a combination of indirect identifiers such as gender, age, education, location etc.

Since the first data protection scandals in the 1990s (Barth-Jones 2012), we have learned that removing or pseudo-anonymising directly identifying attributes (ID's) such as names, addresses and social insurance numbers is generally not sufficient to prevent data protection violations. Beyond masking these direct identifiers, we need to control statistical disclosure risk. This is the risk of intruders using a combination of indirect identifiers such as education level, sex, age, among others, to identify individuals and their health status. The levels of such risks need to be accessed and controlled in the shared datasets to ensure privacy and trust of the individuals contributing their data and to meet scientific integrity requirements.
In the LMIC, the World Bank Group and the International Household Survey Network supported the development of the statistical disclosure control software sdcMicro (Templ, Kowarik, and Meindl 2015) and they recommend it (http://surveys.worldbank.org/sdcmicro).

**Figure 1C. Data flow for deriving an encrypted ID for a target domain (research).** The source ID (1234) can be mapped to the target identifier for example in two ways. A mathematical algorithm is used (left branch) to calculate the target domain ID (6742), or a database query returns the ID. A time stamp is then attached to this ID. ID and timestamp together are then encrypted, e.g. using the public key of the target domain. This assures that no two encrypted IDs for the same patient and the same domain are identical, in this way preventing the unintended linking of records.
APPENDIX 5 – Augmented common data models: An Epi-STACK architecture for COVID-19 epidemiology datasets

Jay Greenfield1,*, Rajini Nagrani2, Meg Sears3, Anna Widyastuti4, Claire C. Austin5; and the RDA-COVID19-WG6

1Data Documentation Initiative, 2Leibniz Institute for Prevention Research and Epidemiology-BIPS, 3Ottawa Hospital Research Institute, 4Hasselt University, 5Environment and Climate Change Canada, 6This work is linked to the Research Data Alliance RDA-COVID19-WG. Epidemiology subWG recommendations and guidelines on data sharing.

*Corresponding author

All views and opinions expressed are those of the co-authors, and do not necessarily reflect the official policy or position of their respective employers, or of any government, agency or organization.

Background
COVID-19 epidemiology crosses many domains. Individuals may be vulnerable due to lifestyle-related factors, health-related behaviours or pre-existing conditions. They may be impacted to varying degrees by the disease, patient care procedures, emergency public health and economic measures, and various sequela. Much of this can be followed via a combination of hospital/clinic surveillance, field-based longitudinal demographic and epidemiological surveillance, and population level indicators. Common data models (CDMs) were first developed to host and support twenty-first century patient care research using EHR data [OHDSI, 2015]. Subsequently, these CDMs were used with EHR data to support emulated clinical trials when clinical trials were impractical [Hernán and Robins, 2016]. Eventually CDMs evolved to include questionnaire data from surveys conducted in the field [Blacketer M, Voss EA, Ryan PB, 2015]. One characteristic of CDMs is that each data element has a semantic annotation. This was necessary with patient care because there are a plethora of electronic health record (EHR) systems, each of which needed to be mapped to the CDM to ensure data comparability across providers. CDMs, however, present a high bar for questionnaires, because they are not always semantically annotated. One benefit of semantically annotating questionnaires in the context of CDMs has been that sections of heterogeneous questionnaires can be mapped to the same CDM data elements, ensuring interoperability across initiatives and locations.

Methods
CDMs are being reviewed in the context of COVID-19. In the present study, we consider the suitability of a CDM in support of a translational research patient care workbench that evaluates the effects of COVID-19 patient care on the clinical course of various patient cohorts.

Other aspirational use cases include an early warning and response system, and a decision support system for public health and economic interventions. In support of these three use cases, we designed a notional Epi-STACK that shapes the Common Data Model so it can host and relate a series of clinical datasets, survey datasets and population indicators over time. This
gives Epi-STACK a time dimension and the ability to produce interim results that are suitable for use in the progress notes that the WHO disseminates over its communication channels (Epi-WIN) and their target audiences.

Results

Figure 1. Patient Care Research Workbench. Proposed workbench to support the construction of emulated trials using big data that crosses the patient care and epidemiological surveillance domains
Figure 2. Epi-Stack. Proposed evidence support system as input to the WHO’s Epi-WIN communication channels targeting various audiences.

Discussion
It remains to be seen how FAIRER (Findable, Accessible, Interoperable, Reusable, Ethical, and Reproducible) CDMs continue when used in support of COVID-19 research. Indeed, there have been several FAIR assessments of a CDM called observational medical outcomes partnership (OMOP) in the context of EHRs. A European Health Data and Evidence Network (EHDEN) initiative based on the FAIR assessment attempts to harmonize 100 million health records [van Bochove K, Vos E, van Winzu A, Kurps J, Moinat M. 2020]. A recent EHDEN initiative is earmarked to use the OMOP CDM for COVID-19 research for registry and cohort data alongside EHRs and hospital information [The EHDEN Consortium. 2020].

Author roles
All authors accept responsibility for the content of the article

References
BIBLIOGRAPHY


“Person Under Investigation (PUI) and Case Report Form.” n.d., 2.


