Sharing COVID-19 Epidemiology Data

Version: 1.0
DOI: 10.15497/rda00049

Authors: RDA COVID-19 Epidemiology Working Group
Co-chair Liaison: Priyanka Pillai
Moderators: Claire Austin, Rajini Nagrani
Published: 4th November 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; guidelines; data sharing

Language: English

License: Attribution 4.0 International (CC BY 4.0)

Citation and Download: RDA COVID-19-Epidemiology WG (2020): Sharing COVID-19 Epidemiology Data, Version 1.0. DOI: 10.15497/rda00049
Research Data Alliance RDA COVID-19

Epidemiology sub-Work Group
Co-moderators: Claire C Austin PhD, Rajini Nagrani PhD, and Gabriel Turinici PhD

SHARING COVID-19 EPIDEMIOLOGY DATA

SUPPORTING OUTPUT
Draft v1.0 (2020-11-04)

CC-BY 4.0 license

Check for the most recent version of this document at: doi.org/10.15497/rda00049
See, also:
- the RDA-COVID19-WG Zotero Library at: doi.org/10.15497/rda00051
- the overarching RDA-COVID19-WG Guidelines and Recommendations across all WGs at:
  https://doi.org/10.15497/rda00052

CONTENTS
Change log .................................................................................................................................................. 2
DATA SHARING IN EPIDEMIOLOGY ........................................................................................................... 3
  1 Focus and Description ............................................................................................................................ 3
  2 Scope ................................................................................................................................................... 3
  3 Policy Recommendations ...................................................................................................................... 3
    3.1 General ............................................................................................................................................... 3
    3.2 Information Technology and Data Management ................................................................................ 3
    3.3 COVID-19 Epidemiological data, analysis, and modeling .................................................................. 4
  3.4 COVID-19 Epidemiological data, analysis, and modeling ................................................................. 4
  4 Guidelines ............................................................................................................................................. 5
    4.1 COVID-19 Population Level Data Sources ......................................................................................... 5
    4.2 Interoperable COVID-19 epidemiological surveillance: Clinical and Population-based instruments .... 6
    4.3 Preservation of individuals’ privacy in shared COVID-19 related data .............................................. 7
    4.4 A Full Spectrum View of the COVID-19 Data Domain: An Epidemiological Data Framework ............. 7
    4.5 Epi-TRACS: A data framework for rapid detection and whole system response for emerging pathogens .. 8
    4.6 COVID-19 Emergency public health and economic measures causal loop: A computable framework .... 8
    4.7 Common Data Models & Full Spectrum Epidemiology: Epi-STACK framework for COVID-19 epidemiology datasets ... 8
ANNEX 1 – COVID-19 Surveillance data and models: Review and analysis ............................................. 10
  Claire C. Austin, Anna Widyastuti, Nada El Jundi, Rajini Nagrani; and the RDA-COVID19-WG
ANNEX 2 – COVID-19 Questionnaires, surveys and item banks: Overview of clinical and population-based instruments ....... 11
  Carsten O. Schmidt, Rajini Nagrani, Christina Stange, Matthias Lübe, Atinkut Zeleke, Guillaume Fabre, Sofiya Koleva, Stefan Sauermann, Jay Greenfield, Claire C Austin; and the RDA-COVID19-WG
ANNEX 3 – Preservation of individuals’ privacy in shared COVID-19 related data ................................... 12
  Stefan Sauermann, Chifundo Kanjala, Matthias Templ; and the RDA-COVID19-WG
  Jay Greenfield, Rajini Nagrani, Meg Sears, Gary Mazzaferro, Claire C. Austin; and the RDA-COVID19-WG
ANNEX 5 – Epi-TRACS: A data framework for rapid detection and whole system response for emerging pathogens such as SARS-CoV-2 virus and the COVID-19 disease that it causes ............................................................ 17
  Jay Greenfield, Henri E.Z. Tonnang, Gary Mazzaferro, Claire C. Austin; and the RDA-COVID19-WG
ANNEX 6 – COVID-19 Emergency public health and economic measures causal loop: Laying the groundwork for a computable framework ............................................................. 23
  Henri E.Z. Tonnang, Jay Greenfield, Gary Mazzaferro, Claire C. Austin; and the RDA-COVID19-WG
ANNEX 7 – Common Data Models & Full Spectrum Epidemiology: An Epi-STACK framework for COVID-19 epidemiology datasets ......................................................................................... 24
  Jay Greenfield, Meg Sears, Rajini Nagrani, Gary Mazzaferro, Anna Widyastuti, Claire C. Austin; and the RDA-COVID19-WG
RDA COVID-19 Epidemiology WG Bibliography .................................................................................. 30-47
CHANGE LOG

<table>
<thead>
<tr>
<th>VERSION</th>
<th>DATE</th>
<th>CHANGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.051</td>
<td>2020-06-08</td>
<td>Fixed version number inconsistencies</td>
</tr>
<tr>
<td>0.051</td>
<td>2020-06-08</td>
<td>Minor edits</td>
</tr>
<tr>
<td>0.051</td>
<td>2020-06-08</td>
<td>Updated Figures</td>
</tr>
<tr>
<td>0.051</td>
<td>2020-06-08</td>
<td>Updated Annexes 1-3, and 5-6</td>
</tr>
<tr>
<td>0.051</td>
<td>2020-06-08</td>
<td>Updated the Bibliography</td>
</tr>
<tr>
<td>0.052</td>
<td>2020-06-15</td>
<td>Updated Data Sharing in Epidemiology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Focus, Description, Scope, Recommendations, Guidelines, and Table 1)</td>
</tr>
<tr>
<td>0.052</td>
<td>2020-06-15</td>
<td>Updated Annex 1 (COVID-19 data sources)</td>
</tr>
<tr>
<td>0.052</td>
<td>2020-06-15</td>
<td>Updated the Bibliography</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated the Bibliography</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Added link to RDA-COVID19-WG Zotero Library</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Added, “Check for the most recent version of the present document at:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>doi.org/10.15497/rda00049”</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Added recommendation for zoonoses in Section 3.3 (Policy Recommendations)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Added guidelines for causal loops in Section 4.6 (Policy Recommendations)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Renumbered the common data model guideline from 4.6 to 4.7</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated Annex 1 (COVID-19 data sources)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated Annex 2 (Questionnaires)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Added zoonoses to Annex 4 (Epi data model)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Split the Epi-Trac paper (Annex 5) into two papers (Epi-TRAC and Causal Loops)</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated Annex 5 (Epi-Trac).</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated Annex 6 (Causal loops).</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Updated Annex 7 (CDM and Epi-STACK).</td>
</tr>
<tr>
<td>0.053</td>
<td>2020-06-28</td>
<td>Miscellaneous minor edits throughout</td>
</tr>
<tr>
<td>0.053b</td>
<td>2020-06-30</td>
<td>Updated link to the new DOI for the final edition (version 1) of the RDA-COVID19-WG Recommendations and guidelines published on 2020-06-30. This is the DOI for the RDA overarching document, not the present supporting output which has not yet been finalized.</td>
</tr>
<tr>
<td>0.053c</td>
<td>2020-06-30</td>
<td>Fixed DOI error and other miscellaneous small errors.</td>
</tr>
<tr>
<td>0.06a</td>
<td>2020-08-31</td>
<td>Updated Annexes 1, 2, 3, 4, 5, 6, and 7.</td>
</tr>
<tr>
<td>0.06a</td>
<td>2020-08-31</td>
<td>Annex #2 (Questionnaires) now available on SSRN preprint server.</td>
</tr>
<tr>
<td>0.06a</td>
<td>2020-08-31</td>
<td>Annex #3 (Privacy) now available on SSRN preprint server.</td>
</tr>
<tr>
<td>0.06a</td>
<td>2020-08-31</td>
<td>Annex #3: Corrected hyperlink to full text.</td>
</tr>
<tr>
<td>0.06b</td>
<td>2020-09-13</td>
<td>Annex #1 (COVID-19 data sources): Updated supplementary tables S1-S8.</td>
</tr>
<tr>
<td>0.06b</td>
<td>2020-09-13</td>
<td>Annex #6 (Causal loops) now available on SSRN preprint server.</td>
</tr>
<tr>
<td>0.06b</td>
<td>2020-09-13</td>
<td>Miscellaneous minor edits.</td>
</tr>
<tr>
<td>0.06c</td>
<td>2020-09-21</td>
<td>Miscellaneous minor edits.</td>
</tr>
<tr>
<td>0.06c</td>
<td>2020-09-21</td>
<td>Annex #1 (Data sources): Updated text and tables</td>
</tr>
<tr>
<td>0.06c</td>
<td>2020-09-21</td>
<td>Annex #1 (Data sources) now available on SSRN preprint server.</td>
</tr>
<tr>
<td>1.0</td>
<td>2020-11-03</td>
<td>Final version</td>
</tr>
</tbody>
</table>
DATA SHARING IN EPIDEMIOLOGY

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.

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.

2 Scope

Epidemiology underpins COVID-19 response strategies and public health measures. The recommendations and guidelines support development of an internationally harmonized specification to enable rapid reporting and integration of epidemiology and related data across domains and between jurisdictions.

The guidelines outline a data driven, coordinated global system that encompasses preparedness, early detection, and rapid response to newly emergent threats such as SARS-CoV-2 virus and the COVID-19 disease that it causes.

Supporting Output

The present supporting output provides supplemental resources, and further develops the global data-driven vision described in the guidelines. This includes a proposed computable framework to support system responses for emerging pathogens. It offers compatible and reliable data models, protocols, and action plans for newly identified threats such as COVID-19.

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

Properly funded state of the art infrastructure is required to support advanced research, as well as the data management and data sharing required for rapid response and collaboration (See Infrastructure Investment). For epidemiology in particular:

1. 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).

2. Rapidly develop standardised tools for aggregating microdata to a harmonised format(s) that can be shared and used while minimising the re-identification risk for individual records.

3. 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. Implement a global system for early detection and rapid response to emerging zoonoses, integrated across systems for reporting human and animal diseases as well as their vectors and geographical distribution (CDC 2020; CGH 2019; eCDC 2019; GLEWS 2006; NASEM 2008, 2009a,b; WHO 2017; WHO, FAO and OIE 2019).

2. Catalogue and document all zoonotic diseases with associated reservoir species and vectors to establish and maintain a global database with potential risks related to humans.

3. 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, hospitalisations 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.
4. Rapidly develop an internationally harmonised specification to enable the export/import/integration of 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.

5. 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 based on industry-standard encryption and cryptographic constructions.


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


9. Harmonise approaches to comparably assess and quantify side-effects of pandemic containment and mitigation measures.

10. Report underlying assumptions and quantify effects of uncertainties on all reported parameters and conclusions for all model predictions etc.

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

4 Guidelines

These guidelines highlight current system challenges and offer solutions to help support a larger framework designed to coordinate and structure the collection and use of COVID-19 related data. Six focus areas, described in the guidelines and supporting output (data sources, instruments, privacy, epidemiological data model, computable framework, and an epi-stack framework), progressively develop a data driven global vision for managing novel biological threats such as COVID-19. We begin with population level data sources that drive the public health strategy and response at all stages of the COVID-19 threat, from emergence through containment, mitigation, and reopening of society. We then survey clinical and population-based instruments that collect data and discuss preservation of individuals’ privacy in shared COVID-19 related data. A full spectrum data model is presented encompassing hospital specific surveillance and electronic health records together with field-based demographic and epidemiological surveillance. We propose Epi-TRACS, a computable framework for emerging pathogen action plans, and an epi-stack that uses the Common Data Model with COVID-19 to integrate clinical data and epidemiological data.

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 organisations 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.

See ANNEX 1 for further details and discussion.
Table 1. COVID-19 population level data sources

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DATA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen Institute for AI</td>
<td>COVID-19 Open Research Dataset (CORD-19)</td>
</tr>
<tr>
<td>Apple Inc.</td>
<td>COVID-19-Mobility Trends Reports</td>
</tr>
<tr>
<td>European Centre for Disease Control</td>
<td>Geographic distribution of COVID-19 cases worldwide</td>
</tr>
<tr>
<td>European Centre for Disease Control</td>
<td>The European Surveillance System (TESSy)</td>
</tr>
<tr>
<td>Institute for Health Metrics and Evaluation (IHME)</td>
<td>Global Health Data Exchange (GHDx)</td>
</tr>
<tr>
<td>Google Inc.</td>
<td>COVID-19 Community Mobility Report</td>
</tr>
<tr>
<td>Johns Hopkins University</td>
<td>COVID19 dataset</td>
</tr>
<tr>
<td>Kieren Healy Rpackage</td>
<td>Rpackage - COVID19 Case and Mortality Time Series</td>
</tr>
<tr>
<td>University of Oxford</td>
<td>COVID19 dataset</td>
</tr>
<tr>
<td>The Atlantic</td>
<td>Tracking Project</td>
</tr>
<tr>
<td>The New York Times</td>
<td>Covid-19 Data in the United States</td>
</tr>
<tr>
<td>U.N.</td>
<td>Humanitarian Data Exchange (HDX)</td>
</tr>
<tr>
<td>U.S. Centre for Disease Control</td>
<td>Cases of COVID19 in the U.S.</td>
</tr>
<tr>
<td>University of Washington</td>
<td>Be Outbreak Prepared</td>
</tr>
<tr>
<td>World Bank</td>
<td>Understanding the Coronavirus (COVID-19) pandemic through data</td>
</tr>
<tr>
<td>World Health Organization (WHO)</td>
<td>Novel Coronavirus (2019-nCoV) situation reports</td>
</tr>
<tr>
<td>Worldometer</td>
<td>COVID19 data</td>
</tr>
</tbody>
</table>

4.2 Interoperable COVID-19 epidemiological surveillance: Clinical and Population-based instruments

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>COUNTRY</th>
<th>QUESTIONNAIRE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CLINICAL</strong></td>
<td></td>
</tr>
<tr>
<td>Australia</td>
<td>NSW Case questionnaire</td>
</tr>
<tr>
<td>Austria</td>
<td>EMS</td>
</tr>
<tr>
<td>Europe</td>
<td>TESSy</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 CoronavirusPerson Under Investigation (PUI) and Case Report Form</td>
</tr>
<tr>
<td><strong>POPULATION BASED</strong></td>
<td></td>
</tr>
<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>
</tbody>
</table>
Table 3. Questionnaire instruments: Resources

<table>
<thead>
<tr>
<th>INITIATIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIH Public Health Emergency and Disaster Research Response (DR2)</td>
</tr>
<tr>
<td>COVID-19 OBSSR Research Tools</td>
</tr>
<tr>
<td>PhenX COVID-19 Toolkit</td>
</tr>
</tbody>
</table>

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 “localisation” information in the question metadata when questions are reused from one survey to the next.

See ANNEX 2 for further details and discussion.

4.3 Preservation of individuals’ privacy in shared COVID-19 related data

Data sharing is essential to improve epidemiological analysis, cross-border pandemic modelling, 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, and clinical condition, 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.

See ANNEX 3 for further details and discussion.

4.4 A Full Spectrum View of the COVID-19 Data Domain: An Epidemiological Data Framework

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. See ANNEX 4 for further details and discussion.

4.5 Epi-TRACS: A data framework for rapid detection and whole system response for emerging pathogens

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 group recommends the creation of a WHO-led EPIdemiological Translational Research Action Coordination System (Epi-TRACS) to add an implementation layer to the existing WHO policies, guidelines, partnerships, and information exchange stack adapted to country-specific contexts. See ANNEX 5 for further details and discussion.

4.6 COVID-19 Emergency public health and economic measures causal loops: A computable framework

Causal loop modeling may be valuable in assessing system sustainability and system resiliency (Bahri 2020; Ricciardi et al. 2020; Wicher 2020). A computable framework in which the actions taken in response to COVID-19 sentinel surveillance can be simulated and assessed both retrospectively and prospectively based on a causal loop diagram may help inform decision making. See ANNEX 6 for further details and discussion.

4.7 Common Data Models and Full Spectrum Epidemiology: An Epi-STACK framework for COVID-19 epidemiology datasets

Common Data Model (CDM)
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 (Hernán et al., 2016). 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.

**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. 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-TRACS early warning and response system, decision support research, and patient care research). It is critical that resiliency be built into the system, from the standpoint of how the system functions under stress. When the degree of complexity and interdependencies increase in human made systems, there is always the risk of collapse if not enough balance is built into the system (both the IT infrastructure, data governance, and the "people" part).

See **ANNEX 7** for further details and discussion.

**Acknowledgments**

The RDA-COVID19-WG would like to acknowledge the following people who have contributed their time, knowledge, and expertise to generate these recommendations and guidelines. Listed alphabetically by last name and ORCID ID:

Claire C. Austin (0000-0001-9138-5986); Marlon L Bayot (0000-0002-5328-150X); Maeve Campman (0000-0002-6613-7144); David Delmail (0000-0003-2836-6496); Cyrille Delpierre (0000-0002-0831-080X); Gayo Diallo (0000-0002-9799-9484); Jay Greenfield (0000-0003-2773-5317); Chifundo Kanjala (0000-0003-0540-8374); Birte Lindstaedt (0000-0002-8251-1597); Gary Mazzaferro (0000-0002-3196-7201); Robin Michelet (0000-0002-5485-607X); Daniel Mietchen (0000-0001-9488-1870); Rajini Nagrani (0000-0002-1708-2319); Carlos Luis Parra-Calderón (0000-0003-2609-575X); Priyanka Pillai (0000-0002-3768-8895); Stefan Sauermann (0000-0003-0824-9989); Carsten Oliver Schmidt (0000-0001-5266-9396); Meg Sears (0000-0002-6987-1694); Henri Tonnang (0000-0002-9424-9186); Gabriel Turinici (0000-0003-2713-006X); and, Anna Widyastuti (0000-0003-2149-935X).
ANNEX 1 – COVID-19 Surveillance data and models: Review and analysis

Claire C. Austin¹,§, Anna Widyastuti²; Nada El Jundi³, Rajini Nagrani³,§; and the RDA-COVID19-WG⁴

¹Environment and Climate Change Canada, ²Hasselt University, Belgium, ³Leibniz Institute for Prevention Research and Epidemiology-BIPS, Germany, ⁴This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

§Corresponding authors: claire.austin@canada.ca and nagrani@leibniz-bips.de


ABSTRACT

Background. Reliable COVID-19 data are critical for understanding the disease and spread of the pandemic, for decision-making, for developing and implementing public health measures, and for tracking the effectiveness of interventions. Currently, however, there is a confusing plethora of publicly available COVID-19 surveillance data resources. Relevant websites are frequently poorly designed making it extraordinarily time-consuming and frustrating to find and extract the relevant information.

Methods. A systemic search of government, official agency, and non-government sources of COVID-19 surveillance and related data, computer code, and forecasting models was conducted.

Results. A comprehensive compendium was built of COVID-19 surveillance data and models having worldwide national coverage, and some sources of particular interest having sub-national coverage. Hyperlinks are provided to download data or computer code from each of the resources. For each resource, a concise description of the data and metadata, including identification of the data sources used to compile the data is provided. The compendium is provided in the supplementary material, organized in nine tables: (1) COVID-19 surveillance datasets and sources; (2) Databases or catalogues of COVID-19 surveillance data; (3) Resources that provide a corpus of COVID-19 related text; (4) Resources that track COVID-19 government responses; (5) R code potentially useful for analysis of COVID-19 data; (6) COVID-19 related data analysis platforms; (7) COVID-19 models; (8) Useful visualizations of COVID-19 data that go beyond the usual ‘dashboards’; and, (9) Commercial sites that showcase their product with a COVID-19 use case. Selected examples of data resources and models are provided in two additional tables in the body of the text.

Conclusion. There is no single source of truth for COVID-19 surveillance data. Government and non-government data were found to be fragmented and difficult to find, access, and interoperate. There is an urgent need to develop a common standard for reporting communicable disease surveillance data based on FAIRER (Findable, Accessible, Interoperable, Reusable, Ethical, and Reproducible) data.

KEYWORDS: COVID-19, data, surveillance, model, sources, computer code, R, FAIRER data.

PLEASE SEE THE SSRN PREPRINT FOR THE FULL TEXT (37 pages)
ANNEX 2 – COVID-19 Questionnaires, surveys, and item banks: Overview of clinical- and population-based instruments

Carsten O. Schmidt¹,*, ⁵ Rajini Nagrani²,*, Christina Stange¹, Matthias Löbe³, Atinkut Zeleke¹, Guillaume Fabre⁴, Sofiya Koleva³, Karine Trudeau⁴, Stefan Sauermann⁵, Jay Greenfield⁶, Claire C Austin⁷; and the RDA-COVID19-WG⁸

¹University Medicine Greifswald, Germany, ²Leibniz Institute for Prevention Research and Epidemiology-BIPS, Germany, ³Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Germany, ⁴Maelstrom Research Institute, McGill University, Canada, ⁵UAS Technikum Wien, Austria, ⁶Data Documentation Initiative, USA, ⁷Environment and Climate Change Canada; ⁸This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

* Co first authors: Schmidt and Nagrani

§ Corresponding author: carsten.schmidt@uni-greifswald.de


ABSTRACT

New COVID-19 related instruments are being rapidly developed around the world to collect patient- and population-based information. Heterogeneity between instruments limits comparability of results. The present study provides an overview of instruments and resources. We scoped the content domain on a selection of instruments using the Maelstrom taxonomy. Content of the instruments varied widely, from proximal measures (e.g., clinical symptoms, comorbidities, etc.) to distal measures such as political attitudes. We recommend that researchers reuse existing instruments to the greatest extent possible, and that they make results openly available in machine-readable format to facilitate reuse and maximize comparability of results across studies and countries.

Keywords: COVID-19, SARS-CoV-2, questionnaire, survey, collection tool, instrument, epidemiology, clinical, population, symptoms, interoperability, taxonomy, semantic annotation.

PLEASE SEE THE SSRN PREPRINT FOR THE FULL TEXT (9 pages)
ANNEX 3 – Preservation of individuals’ privacy in shared COVID-19 related data

Stefan Sauermann¹, Chifundo Kanjala², Matthias Templ³,*, Claire C. Austin⁴; and the RDA-COVID19-WG⁴

¹UAS Technikum Wien, ²London School of Hygiene and Tropical Medicine, ³Zurich University of Applied Sciences, ⁴Environment and Climate Change Canada, ⁵This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing; and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

*Corresponding author: chifundo.kanjala@lshtm.ac.uk


ABSTRACT
This paper provides insight into how restricted data can be incorporated in an open-be-default-by-design digital infrastructure for scientific data. We focus, in particular, on the ethical component of FAIRER (Findable, Accessible, Interoperable, Ethical, and Reproducible) data, and the pseudo-anonymization and anonymization of COVID-19 datasets to protect personally identifiable information (PII). First, we consider the need for the customisation of the existing privacy preservation techniques in the context of rapid production, integration, sharing and analysis of COVID-19 data. Second, the methods for the pseudo-anonymization of direct identification variables are discussed. We also discuss different pseudo-IDs of the same person for multi-domain and multi-organization. Essentially, pseudo-anonymization and its encrypted domain specific IDs are used to successfully match data later, if required and permitted, as well as to restore the true ID (and authenticity) in individual cases of a patient’s clarification. Third, we discuss application of statistical disclosure control (SDC) techniques of COVID-19 disease data. To assess and limit the risk of re-identification of individual persons in COVID-19 datasets (that are often enriched with other covariates like age, gender, nationality, etc.) to acceptable levels, the risk of successful re-identification by a combination of attribute values must be assessed and controlled. This is done using statistical disclosure control for anonymization of data. Lastly, we discuss the limitations of the proposed techniques and provide general guidelines on using disclosure risks to decide on appropriate modes for data sharing to preserve the privacy of the individuals in the datasets.

KEYWORDS: Pseudo-anonymization, statistical disclosure control, data anonymization, data sharing, privacy, personally identifiable information, PII, COVID-19, open science.

PLEASE SEE THE SSRN PREPRINT FOR THE FULL TEXT (13 pages)
DISCUSSION PAPER

ANNEX 4 – Full Spectrum View of the COVID-19 data domain: An Epidemiological Data Framework

Jay Greenfield¹, §, Rajini Nagrani², Meg Sears³, Gary Mazzaferro⁴, Claire C Austin⁵, §; and the RDA-COVID19-WG⁶

¹Data Documentation Initiative, USA, ²Leibniz Institute for Prevention Research and Epidemiology-BIPS, ³Ottawa Hospital Research Institute, ⁴Consultant, ⁵Environment and Climate Change Canada, ⁶This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

§ Corresponding authors: nightcleaner@gmail.com and claire.austin@canada.ca


ABSTRACT

We propose an epidemiological surveillance conceptual data framework to guide construction of a Common Data Model and database schema for COVID-19 epidemiological research. The framework encompasses hospital specific surveillance in line with the WHO COVID-19 core and rapid case report forms, electronic health records, and field-based COVID-19 demographic and epidemiological surveillance.

KEYWORDS: COVID-19, epidemiology, disease milestones, contacts, personal risk factors.

BACKGROUND

This is the first paper in a series of four discussion papers that progressively build a proposed Epi-STACK framework to support full spectrum COVID-19 epidemiology. Based on the traditional, deterministic Susceptibility, Exposure, Infection, Recovery (SEIR) model, the present paper introduces the concept of a full spectrum epidemiological surveillance framework. The second paper in the series proposes a conceptual early warning and response framework for development of an, “Epidemiology Translational Research Action Coordinated System (Epi-TRACS). The third paper proposes the use of systems thinking and causal loop diagrams (CLD) to better understand multiple interrelated factors and impacts. The last paper in the series proposes an Epi-STACK framework for development of COVID-19 datasets based on the Common Data Model to support full spectrum epidemiology to support public health recommendations. This series of four papers follows upon a review and analysis of COVID-19 population level data sources (Austin et al. 2020), an overview of COVID-19 clinical and population based
questionnaire and survey instruments (Schmidt et al. 2020), and explores how restricted data can be incorporated in an open-be-default-by-design digital infrastructure for scientific data with preservation of individuals’ privacy in the COVID-19 context (Sauermann et al. 2020).

The full spectrum COVID-19 domain framework¹ accounts for the individual’s entire experience, including susceptibility, exposure, disease severity, infection, treatment, sequelae, and either death or recovery (Pesquita et al. 2014; WHO 2020 Apr 23; WHO 2020 Mar 24). Susceptibility includes person-level risk factors, public health measures, and individuals’ compliance with public health orders and recommendations.

A FULL SPECTRUM EPIDEMIOLOGY FRAMEWORK

We propose development of an augmented full spectrum data framework based on the Susceptibility, Exposure, Infection, Recovery (SEIR) model (Sun and Kang 2020). The framework would take into account public health measures (Giordano et al. 2020) and differences in the availability of resources between High Income Countries (HIC) and Low and Middle Income Countries (LMIC). In HICs, the SEIR model may be more informed by the hospital experience, while in LMICs it may be community-based demographic and epidemiological surveillance that carry greater influence (Wang et al. 2020). Some LMICs also benefit more directly from lessons learned from prior experience with Ebola and HIV AIDS. A COVID-19 epidemiological surveillance data framework would incorporate clinical, community, and epidemiological surveillance data (Figure 1).

Figure 2 breaks out the framework into three components: (A) disease milestones; (B) contacts; and, (C) personal risk factors, and identifies their provenance. Disease milestones and the personal risk factors components would be based on the WHO Global Influenza Surveillance and Response System (GISRS 2020a), COVID-19 sentinel surveillance (WHO 2020b), Wellcome Trust Longitudinal Populations Strategy (Wellcome Trust 2017), WHO COVID-19 core version and rapid version case report forms (CRF) (WHO 2020b; WHO 2020c), and the Wellcome Trust COVID-19 LPS HIC and LMIC Questionnaires (Wellcome Trust COVID-19 LPS Questionnaires).

¹ In ontology engineering, a domain model is a formal representation of a knowledge domain with concepts, roles, data types, individuals, and rules, typically grounded in a description logic.
Figure 1: COVID-19 epidemiological surveillance data framework concept

Figure 2: Overview of the proposed COVID-19 epidemiological surveillance data framework
DISCUSSION

The proposed epidemiological surveillance data framework would inform development of an extensible database schema. See, for example, schema.org (2020) and HL7 FHIR (2020). The proposed framework is longitudinal in scope, and the extensible schema would need to be longitudinal as well to account for such things as repeat visits to health care facilities (e.g., hospitals, clinics, and nursing homes), changed diagnoses, multiple diagnostic tests, and repeat encounters with field workers. New contact tracing tools will need to be developed for large scale encounters where traditional methods fail. The personal risk factors component would be refined as greater understanding is gained about COVID-19 disease, how it spreads, and the effectiveness of different public health measures.

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

Funding: None.
Conflict of interest: None to declare.

REFERENCES


DISCUSSION PAPER

ANNEX 5 – Epi-TRACS: Rapid detection and whole system response for emerging pathogens such as SARS-CoV-2 virus and the COVID-19 disease that it causes

Jay Greenfield¹, §, Henri E.Z. Tonnang², Gary Mazzaferro³, Claire C. Austin⁴, §; and the RDA-COVID19-WG⁵

¹Data Documentation Initiative, USA, ²International Center of Insect Physiology and Ecology (ICIPE), Kenya, ³Consultant, USA, ⁴Environment and Climate Change Canada, ⁵This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG. Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.  

§ Corresponding authors: nightcleaner@gmail.com and claire.austin@canada.ca

CITE AS: Jay Greenfield, Henri E.Z. Tonnang, Gary Mazzaferro, Claire C. Austin; and the RDA-COVID19-WG. (2020). Epi-TRACS: Rapid detection and whole system response for emerging pathogens such as SARS-CoV-2 virus and the COVID-19 disease that it causes. [IN: Sharing COVID-19 epidemiology data, Research Data Alliance RDA COVID-19 Epidemiology WG]. https://doi.org/10.15497/rda00049

ABSTRACT

Despite repeated warnings and recommendations to prepare for emerging pathogens that could result in pandemics, the world was not prepared to respond to the fast moving threat of COVID-19. The present paper proposes a conceptual early warning and response framework for “Epidemiology Translational Research Action Coordinated System” (Epi-TRACS).

KEYWORDS: COVID-19, epidemiology, surveillance, economy, causal analysis

BACKGROUND

COVID-19 threat detection was slow and ineffective, resulting in rapid development of a pandemic. Countries around the world have implemented a disparate series of public health measures in attempting to suppress and mitigate spread of the disease. The world was not prepared to respond to a novel zoonose that spreads with the tempo and severity of COVID-19. The result has been widespread lockdowns which have had serious socioeconomic consequences for both High Income Countries (HIC) and for Low and Middle Income Countries (LMIC) (Gates 2020; Bai et al. 2020).

Previous outbreaks such as H5N1 in 1997, SARS-1 in 2003, H7N2 in 2004, MERS-CoV in 2012, Ebola in 2014, and others, should have been a wake-up call. Failure to heed warnings and recommendations has had expensive and tragic consequences (NASEM 2008, 2009a,b, 2010, 2015, 2018, 2020a,b,c).
Over the past 20 years, before COVID-19, it is estimated that zoonotic diseases cost $100US billion in economic losses, and killed two million people every year - mostly in LMICs (UNEP and ILRI 2020). It has been estimated that the COVID-19 pandemic will cost $9 trillion over the next few years (UNEP and ILRI 2020).

In 2008, the National Academies of Sciences, Engineering, and Medicine (NASEM) conducted a workshop on achieving sustainable global capacity for surveillance and response to emerging diseases of zoonotic origin:

“These newly identified diseases have emerged primarily as a result of significant changes in human activity, including population growth, increased demand for animal protein, increased wealth and rapid travel by people and their animals, changes to the environment, and human encroachment on farm land and previously undisturbed wildlife habitats. Other pathogens could follow a similar pathway. ... It is very important for policy makers to understand the kind of surveillance and action that will be needed to protect the public and the benefits they provide, and it is up to the scientific and public health community to make this case. ... The current status of global surveillance systems in both human and animal populations and the strength of the veterinary health systems are insufficient to preempt a pandemic or to handle an emerging one. “What’s needed is a new paradigm, a means for tapping expertise from all sectors, and thinking in a broadly preventive way, to reform animal husbandry and alter the ways people and industries interact with domestic animals and wildlife. The interactions among many factors—from rapid mass transportation to increased consumption of animal protein to wilderness encroachment—have intensified the threat posed by zoonotic diseases. The global community involved with disease surveillance and coordination will be needed to confront this challenge. ... The challenge for controlling emerging diseases with such complex etiology is that no single agency has either the mandate or the capacity to address the entire landscape of zoonotic disease. ... The immediate challenge is for the world to support disease monitoring in those areas that need it, while the longer term challenge is to develop greater political will to support a truly global approach to surveillance. ... The most promising approach to sustainable global disease surveillance and response is international collaboration, but large, international organizations are only part of the answer. Sustainable, collaborative work must be based close to where the problems are” (NASEM 2008).

NASEM (2008) noted that surveillance, defined by the WHO as “the systematic ongoing collection, collation, and analysis of data for public health purposes and the timely dissemination of public health information for assessment and public health response as necessary,” had been developing in an ad hoc manner. Important organizations conducting surveillance included the Food and Agriculture Organization of the United Nations (FAO), World Organization for Animal Health (OIE), and the World Health Organization (WHO), in addition to surveillance conducted by countries. In 2006, the joint FAO-OIE-WHO Global Early Warning System (GLEWS) was established to assist in early warning, prevention and control of animal disease threats, including zoonoses (WHO 2006). In 2013, the organization
evolved to GLEWS+ to, “to advance from reactive to proactive preparedness and prevention, through joint risk assessment for targeted and timely action” (GLEWS 2013).

LMICs and HICs rely on the WHO Global Influenza Surveillance and Response System (GISRS) (WHO, 2020a). HICs also have some form of additional early response system. GISRS laboratories have become COVID-19 testing centres in many countries. Syndromic and sentinel surveillance have monitored community transmission and geographic spread of influenza-like illness (ILI), acute respiratory infection (ARI), and severe acute respiratory infection (SARI). Continued vigilance is needed to detect the emergence of novel zoonotic viruses affecting humans (WHO 2020b).

Presently, there is a need to standardize and harmonize COVID-19 data collection and reporting across jurisdictions within existing surveillance systems (WHO 2020b).

EARLY WARNING AND RESPONSE DATA SYSTEM: EPI-TRACS

We propose a conceptual early warning and response Epidemiology Translational Research Action Coordinated System (Epi-TRACS) inspired by Activity Based Intelligence (ABI) (Figure 1). Quinn (2012) describes ABI as:

Activity Based Intelligence is an analysis methodology which rapidly integrates data from multiple traditional data sources and other sources around the interactions of people, events and activities, in order to discover relevant patterns, determine and identify change, and characterize those patterns to drive collection and create decision advantage.

Attwood (2015) emphasises that “the enterprise must embrace the opportunities inherent to Big Data while also driving toward a unified intelligence strategy.” He notes that, “the primary strategy thus far has been acquisition based, looking to industry and research and development organizations to provide the next best tool and software, rather than addressing the more existential requirement of advancing analytical tradecraft and transforming antiquated intelligence analysis and processing methods.”

Atwood's (2015) further emphasises that, “to truly revolutionize and fundamentally change from an individual exploitation process to analysis-based tradecraft, the enterprise needs to harness the potential of Big Data, replacing the methodology of individually exploited pieces of data with an ABI approach. This will enable analysts to focus on hard problems with critical timelines as well as normal day-to-day production activities.” This is also true in the epidemiology domain, most acutely evident in the case of emerging pathogens and pandemics. In epidemiology, also, the sharp incline in the amount of data, recent information technology (IT) advances, and the ABI methodology impel significant changes within the traditional intelligence production model of PCPAD (planning and direction, collection, processing and exploitation, analysis and production, and dissemination) (Atwood (2015).
Epi-TRACS would use coordinated, rapid response teams for early identification and evaluation of pathogens. To be effective, Epi-TRACS would need to be integrated into existing international surveillance systems across countries and domains. The Epi-TRACS data framework includes sentinel surveillance, public health disease management tools, inclusive growth and recovery tools. Full spectrum epidemiology and the Common Data Model components shown in Figure 1 are discussed elsewhere (Greenfield et al. 2020a,b). The principles of findable, accessible, interoperable, reusable, ethical, and reproducible (FAIRER) data principles and transparency built into Epi-TRACS would contribute to the improving development of decisions and action plans targeting public health and economic outcomes.
AUTHOR ROLES

All authors accept responsibility for the content of the article
Writing: JG, CCA. Bibliographic review and analysis: JG, CCA, MC. Visualization: JG, HT, CCA.

Funding: None.
Conflict of interest: None to declare.

REFERENCES

Atwood CP. Activity-Based Intelligence: Revolutionizing Military Intelligence Analysis. Joint Force Quarterly. 2015; 77 (2nd Quarter, April 2015)
https://pdfs.semanticscholar.org/b96e/d4e1bcd62fc79eed470db58398d8fbbe514b.pdf?_ga=2.164032895.1918407571.1598743582-256986716.1598743582


https://doi.org/10.17226/21690


DISCUSSION PAPER

ANNEX 6 – COVID-19 Emergency public health and economic measures causal loop: Laying the groundwork for a computable framework

Henri E.Z. Tonnang¹, Jay Greenfield², Gary Mazzaferro³, Claire C. Austin⁴, §; and the RDA-COVID19-WG⁵

¹International Center of Insect Physiology and Ecology (ICIPE), Kenya, ²Data Documentation Initiative, USA, ³Consultant, USA, ⁴Environment and Climate Change Canada, ⁵This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

§ Corresponding authors: htonnang@icipe.org and claire.austin@canada.ca


ABSTRACT

COVID-19 infections follow dynamic patterns across society. We propose using a “systems thinking” approach to better understand multiple interrelated factors and impacts. Building upon the work of others, we have extended the use of Causal Loop Diagrams (CLD) for whole system qualitative analysis of the linkages and interrelationships between COVID-19 contagion, healthcare, and economic components. The approach used is generic and can easily be applied to both developing and developed countries.

KEYWORDS: Contagion, healthcare, economy, holistic, system thinking, reinforcing and balancing, causal loop diagram (CLD), systems research.

PLEASE SEE THE SSRN PREPRINT FOR THE FULL TEXT (8 pages)
DISCUSSION PAPER

ANNEX 7 – Common Data Models and Full Spectrum Epidemiology: An Epi-STACK framework for COVID-19 epidemiology datasets

Jay Greenfield¹, §, Meg Sears², Rajini Nagrani³, Gary Mazzaferro⁴, Anna Widyastuti⁵, Claire C. Austin⁶, §; and the RDA-COVID19-WG⁷

¹Data Documentation Initiative,²Ottawa Hospital Research Institute, ³Leibniz Institute for Prevention Research and Epidemiology-BIPS, ⁴Consultant, ⁵This work was developed as part of the Research Data Alliance RDA-COVID19-WG Recommendations and guidelines on data sharing, and the RDA COVID-19 Epidemiology WG Data sharing in epidemiology, and we acknowledge the support provided by the RDA community and structures.

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.

§ Corresponding authors: nightcleaner@gmail.com and claire.austin@canada.ca


ABSTRACT

We propose an Epi-STACK framework for hosting COVID-19 datasets compatible with the Common Data Model (CDM). The framework supports full spectrum COVID-19 research, epidemiological and demographic surveillance, clinical care data, and emulated trials.

KEYWORDS: Common Data Model, CDM, epidemiology

BACKGROUND

The World Health Organization (WHO) established an Information Network for Epidemics (EPI-WIN) following the outbreak of COVID-19 disease (WHO 2020). The present paper is the last in a series of four discussion papers leading to the proposal of an Epi-STACK framework based on Common Data Models (CDM). We previously proposed a full-spectrum epidemiology data framework (Greenfield et al. 2020a), Epi-TRACCS early warning and response system (Greenfield et al. 2020b), and a computable framework based on systems analysis (Tonnang et al. 2020). The present paper provides an activity-based intelligence framework that could help support development of WHO recommendations and guidelines that are communicated and amplified by the WHO EPI-WIN network.

Clinical and observational research are often conducted as stand-alone studies with little integration between them even when conducted in the same population. The purpose for which clinical data are collected may not meet the data requirements of epidemiology studies. In addition, medical personnel may compromise the quality of data collected when the data are not perceived to be relevant to treatment.
THE COMMON DATA MODEL

Common Data Models (CDMs) have been developed to host and support twenty-first century patient care research using Electronic Health Record (EHR) data. CDMs require standardized/consistent data to facilitate the exchange, pooling, sharing, or storing of data from multiple sources. Within the last decade, several CDMs have been developed collaboratively, and have risen to the level of de facto standards for clinical research data (Garza et al. 2016). CDMs may also enable the exploration of observational, quasi-experimental, and experimental clinical studies using healthcare data (Garza et al. 2016). Quasi-experimental designs include the use of EHR data to support emulated clinical trials when clinical trials are impractical (Hernán and Robins, 2016). CDMs support multiple data sources and have evolved to include questionnaire data from field surveys (Blacketer et al. 2015).

Combined data from EHR and survey questionnaires is integral to an epidemiology intelligence framework incorporating person-level clinical and community data to evaluate public health measures. Such an intelligence framework supports cross-domain full spectrum epidemiology (Atwood, 2015).

In parallel with development of CDM approaches, efforts have been made to streamline Case Report Forms (CRFs) and population-based questionnaires (Garza et al. 2016). However, little has been done to integrate the two. We therefore propose the use of the CDM to integrate CRFs and COVID-19 epidemiology data in an Epi-STACK.

In support of Health Care risk assessment research, inputs to the Common Data Model may include health data and community data that take the form of epidemiological surveillance questionnaires administered repeatedly over time (Figure 1).

Figure 1. Full spectrum translational research enabled by the Common Data Model. The Common Data Model can be a full spectrum provider for emulated trials. [SOURCE: Book of OHDSI, Chapter 4 (Common Data Model), C. Blacketer, 2020. Creative Commons CC0 v1.0.]
The Common Data Model includes schemas in support of clinical care data and questionnaires, including standardized metadata. Standardized metadata in turn makes full spectrum epidemiology FAIRER (Findable, Accessible, Interoperable, Reusable, Ethical and Reproducible). In fact, several FAIR assessments of a CDM called the Observational Medical Outcomes Partnership (OMOP) have been undertaken (OHDSI 2015). The European Health Data and Evidence Network (EHDEN) initiative tested OMOP for FAIR in a task that required OMOP to harmonize 100 million health records gathered from many sources (van Bochove et al. 2020). A more recent EHDEN initiative has tasked OMOP to go full spectrum in COVID-19 research by combining registry and cohort data alongside EHRs and hospital information (The EHDEN Consortium 2020).

The CDM also supports a risk assessment in emulated trials (Figures 1 and 2), where the impact of various disease management and treatments taken separately or together may be assessed using EHR and/or claims data alongside exposure and/or other confounding/stratification variables (Gershman et al. 2018).

Figure 2. COVID-19 Risk Assessment for Health Care and Population Health.
EPI-STACK

An activity-based intelligence framework was adapted to create Epi-STACK to support cross-domain, full spectrum COVID-19 epidemiology (Figure 3). Epi-STACK may support many types of tools and analytical models surrounding pandemic response. Emulated trials can be explored to discern associations and causal relationships, as illustrated in Figure 2. The COVID-19 Risk Assessment Research Framework is shown in Figure 3.

Figure 3. Epi-STACK: A framework inspired by Activity Based Intelligence to support full spectrum COVID-19 epidemiology.

Full spectrum epidemiology considers multiple data types disease surveillance sources together with the public health and economic development that mitigate spread and severity of disease in a population (Figure 2). The proposed full spectrum COVID-19 epidemiology (Figure 2) would be hosted by a CDM. One of the many research scenarios that Epi-STACK could support would be a Risk Assessment Research Framework for COVID-19 Health Care using causal analysis in quasi-experimental designs (Figure 3).
Epi-STACK and EPI-WIN

The CDM is the starting point of a customized workflow to host full spectrum epidemiology (Figure 4).

![Figure 4. Epi-STACK showing three use cases.](image)

Risk assessment using emulated trials is one use case for the Common Data Model. Other use cases could include an early warning and response system, decision support systems for public health and economic interventions (Tonnang et al. 2020), as well as Epi-TRACS (Greenfield et al. 2020). These use cases define our proposed Full Spectrum Epidemiology, which in turn gives shape to the CDM. All of these use cases require a time series of clinical datasets, survey datasets, and population indicators which make Epi-STACK fit for purpose. Epi-STACK has the ability over time to produce a series of interim results that WHO Epi-WIN would be able to channel to a wide range of target audiences (Figure 4).

**Author roles**

All authors accept responsibility for the content of the article

**Conceptualization:** GM MS RN CCA. **Methodology:** GM HT. **Investigation:** MC. **Formal analysis:** CCA. **Validation:** CCA MC. **Writing:** JG MS CCA. **Bibliographic review and analysis:** MC AW. **Visualization:** JG

**Funding:** None.

**Conflict of interest:** None to declare.
REFERENCES

Atwood CP. Activity-Based Intelligence: Revolutionizing Military Intelligence Analysis. *Joint Force Quarterly*. 2015; 77 (2nd Quarter, April 2015)

Blacketer M, Voss EA, Ryan PB. Applying the OMOP Common Data Model to Survey Data [Webpage]. 2015

  Available from: https://www.bmj.com/content/369/bmj.m1985


van Bochove K, Vos E, van Winzu A, Kurps J, Moinat M. Implementing FAIR in OHDSI and EHDEN. 

  https://www.who.int/teams/risk-communication


https://doi.org/10.20944/preprints202005.0033.v1


GESIS Panel Team. (2020). GESIS Panel Special Survey on the Coronavirus SARS-CoV-2 Outbreak in Germany. [Data set]. GESIS Data Archive. https://doi.org/10.4232/1.13520


NIST PWG. (2019a). *Big Data Interoperability Framework: Volume 1, Definitions* (NIST SP 1500-1r2; p. NIST SP 1500-1r2, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-1r2](https://doi.org/10.6028/NIST.SP.1500-1r2)

NIST PWG. (2019b). *Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements* (NIST SP 1500-3r2; p. NIST SP 1500-3r2, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-3r2](https://doi.org/10.6028/NIST.SP.1500-3r2)

NIST PWG. (2019c). *Big Data Interoperability Framework: Volume 4, Security and Privacy* (NIST SP 1500-4r2; p. NIST SP 1500-4r2, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-4r2](https://doi.org/10.6028/NIST.SP.1500-4r2)

NIST PWG. (2019d). *Big Data Interoperability Framework: Volume 6, Reference Architecture* (NIST SP 1500-6r2; p. NIST SP 1500-6r2, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-6r2](https://doi.org/10.6028/NIST.SP.1500-6r2)

NIST PWG. (2019e). *Big Data Interoperability Framework: Volume 7, Standards Roadmap* (NIST SP 1500-7r2; p. NIST SP 1500-7r2, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-7r2](https://doi.org/10.6028/NIST.SP.1500-7r2)

NIST PWG. (2019f). *Big Data Interoperability Framework: Volume 8, Reference Architecture Interfaces* (NIST SP 1500-9r1; p. NIST SP 1500-9r1, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-9r1](https://doi.org/10.6028/NIST.SP.1500-9r1)

NIST PWG. (2019g). *Big Data Interoperability Framework: Volume 9, Adoption and Modernization* (NIST SP 1500-10r1; p. NIST SP 1500-10r1, version 3 (final)). National Institute of Standards and Technology, Big Data Public Working Group. [https://doi.org/10.6028/NIST.SP.1500-10r1](https://doi.org/10.6028/NIST.SP.1500-10r1)


RDA-COVID19 Zotero WG. (2020, June 5). *RDA-COVID19 WG Zotero Library*. RDA. [https://doi.org/10.15497/rda00051](https://doi.org/10.15497/rda00051)


