We Can Do Better: Lessons Learned on Data Sharing in COVID-19 Pandemic Can Inform Future Outbreak Preparedness and Response

Jonathan LoTempio, D'Andre Spencer, Rebecca Yarvitz, Arthur Delot Vilain, Eric Vilain, and Emmanuèle Délot

The COVID-19 pandemic will remain a critical issue until a safe and effective vaccine is in global use. A strong international network exists for the systematic collection and sharing of influenza genome sequence data, which has proven extensible to COVID-19.¹ However, the robust demographic and clinical data needed to understand the progression of COVID-19 within individuals and across populations are collected by an array of local, regional, federal, and/or national agencies, with country-specific, often overlapping mandates. The networks tasked with transmitting descriptive, disaggregated data have not done so in a standardized manner; most data are made available in variable, incompatible forms, and there is no central, global hub.

At the same time, data have been released and analyzed on Twitter, in blogs, and by news outlets before undergoing scholarly peer review in the race to learn about the emerging virus. Many journals have lowered their paywalls on COVID-19 papers in accordance with WHO guidelines for sharing data in public health emergencies. This has led to unprecedented access to knowledge for anyone,

Information about the authors can be found at the end of the article.
anywhere. Furthermore, in peer-reviewed journals, more papers that include the term “coronavirus” have been published in just the past six months than in the previous 70 years (Figure 1). Sorting through these papers to inform evidence-based policy is daunting. This wealth of new knowledge can be a curse: scientists and policymakers drown in the open sea of literature; unsure which papers are life rafts and which are flotsam.

In short, there is a need for improved international data sharing policies to support outbreak response. Previous successes and challenges in data sharing are instructive. We probed the national-level reporting on cases between April 1 and July 1, 2020 (evidence), efforts to share the pathogen’s genome sequence (science), and existing tools for assessing and disseminating publications (policy). Based on these observations, we suggest several structural improvements to the systems that facilitate knowledge sharing.

**Figure 1**

**Figure 1. Publications with the term “coronavirus” archived in PubMed over time.** Since the start of the pandemic, more papers have been published on coronaviruses than all previous years combined. This fact underscores the need for expert guidance in combing through and making sense of the literature. The emergence of the two previous human coronavirus outbreaks were followed by noticeable but much smaller increases in publications. The term coronavirus was published for the first time in 1968 (see “News and Views” in *Nature*, vol. 220, no. 5168, November 16, 1968, p. 650). PubMed indexes anterior reports of infections later demonstrated to be due to the same type of virus, likely via the use of keywords (the early publications are not accessible online via PubMed). Due to the size of the search query, the data underlying this figure were generated with Entrez Direct command line search.
Observations

Evidence: Heterogenous, internationally non-interoperable COVID-19 case reporting

We focused on the data reported by public health agencies of the fifteen nations with the highest COVID-19 burden.² As of April 23, 2020, this list included the United States (US), Spain, Italy, France, Germany, France, the United Kingdom (UK), Turkey, Iran, China, Russia, Brazil, Belgium, Canada, the Netherlands, and Switzerland, which represented ~75% of the reported global cases. Features of the data reported by these agencies on April 23 can be examined in the Annex, along with a table updated as of July 1, 2020.³

Many languages of data reports: The many tongues of the global village are reflected in the ten unique de facto or national languages in these fifteen countries. While summary data are reported in English by all nations except Iran, China, Brazil, and Russia, more detailed reports, where available, are typically in official languages. We used Google Translate to compare the official websites.

Heterogeneous data format: All fifteen agencies reported data in plain text, HTML, or PDF; eleven offered an interactive web-based data dashboard; and seven had comma-separated (CSV) data for download. Between April 1 and April 23, the US agency added a CSV download function and the Turkish agency switched from a CSV format to a Google product. None of these CSV files are compatible, and no codebook was appended for decoding. There is little to no documentation which records the locations where old data are archived or which URLs are stable. The diversity of these fundamental methods of distribution presents great challenges to accessing the data.

Data aggregation status: All agencies recorded cases and deaths of citizens due to COVID-19, while only nine nations reported the number of tests conducted within their borders. The number of tests performed can help shape evidence-based policy by offering a crude estimate of infectivity.

Only eight agencies reported on age and sex, yet both variables are known to have a strong predictive value on health outcomes. Age was reported categorically with different, incompatible age groups across agencies, thus limiting research potential.⁴,⁵ No agencies included socioeconomic status in their reports. Race/ethnicity/ancestry was largely ignored, presented only by the US after April 17.
Clinical aspects of the disease, such as comorbidities, symptoms, or admission to intensive care, were also unevenly reported. No agencies reported reinfection. Accounting for each of these factors is of key importance to the professionals looking to track, treat, cure, and prevent this disease.

**Science:** A robust system already exists for the rapid sharing of viral genome sequence data

The Bermuda Accord,\(^6\) outlined in 1996 at an International Human Genome Project meeting, was a progenitor of rapid data release protocols, with the express purpose of the betterment of society. This statement informed the HapMap data-sharing policies, which still underpin human genomics today.\(^7\) In turn, HapMap data-sharing policies inspired the founders of the 2006 Global Initiative on Sharing Avian Influenza Data (GISAID).\(^8\) GISAID further required that data users not only give credit to data submitters, but make maximum efforts to work with and include them in joint analyses on viral sequence data, further tipping the scales in favor of collaboration. This mandated sharing of not only data, but the benefits of research, has resulted in a paradigm shift which helps to put contributors from higher or lower resource settings on the same footing.

More than a decade and a name change since its conception, the nearly invisible infrastructure of the Global Initiative on Sharing All Influenza Data (not just avian influenza anymore), influences both research and actionable policy, from the selection of yearly flu vaccines and the close monitoring of emerging strains to the less tangible but equally important dimensions of diplomacy, including conflict resolution and trust building.\(^9\) During the 2009 H1N1 pandemic, there was only a 10-day period between the first human case identified in the US,\(^10\) and the deposition of a full influenza genome sequence in GISAID.\(^11\) The GISAID team ensured that the repository was globally accessible by leveraging team members across multiple time zones, allowing for round the clock work.\(^12\) Since then, the system has dealt with both light and heavy flu seasons, accommodating the yearly influx of data and demand for access to those data.

Through these efforts, GISAID has demonstrated that career advancement and global cooperation are compatible, with little negative impact on researchers’ ability to make a name for themselves, a perennial concern in the open data world. By changing the default from closely held data and few-author publications to openness and rapid, multi-author publications, GISAID has helped influenza research blossom.

GISAID has also shown remarkable robustness in the face of the COVID-19 pandemic. Since the first two SARS-CoV-2 genome sequence submissions on January 10, 2020, more than 57,000 have followed (through July 1, 2020), deposited
from every continent except Antarctica and most countries. These files are high-quality, deeply sequenced genomes with few errors, not just the results of nasopharyngeal swab samples. For comparison, the National Center for Biotechnology Information (NCBI, US National Institutes of Health) Genbank and European Nucleotide Archive only contain ~10,000 SARS-CoV-2 sequences each. The Chinese National Genomic Data Center’s Genbank links directly to GISAID.

The GISAID epiCoV database includes metadata which are largely genomics-focused: sequence of the virus in fasta (plain text) format; and sequencing platform and analytic tools used to generate the sequence. Metadata are organized by information about the virus clade, sample, institute, and submitter. These can be further annotated with team member names, geographic location of sample, sampling methodology, the age and sex of the case, and any treatment, if known. However, filtering the database to only include sequences from patients of known clinical status reduces the number to approximately 3,300. While each viral sequence submission includes a unique identifier, detail on whether multiple sequences from one individual are deposited is not readily available. Given the tens of thousands of available sequences, the majority of the metadata in the database is incomplete – just as with the public national dashboards described above.

Policy: an explosion of academic manuscripts underscores the need for curation and summary

The products of academic research are key for informing the design of new studies and policy. These products include two broad categories: preprints and journal publications. Preprints have little or no review prior to being posted, though ethical statements and declaration of competing interests are usually required for the protection of human and/or animal subjects. Journal publications undergo editorial review by the journal or platform, and typically peer review as well, where professionals in the field are asked to assess the scientific and scholarly merits of papers submitted to the journal.

In the 70 years between 1949 and 2019, approximately 14,000 papers referencing coronavirus of any kind were published and indexed by PubMed, a database maintained by NCBI (Figure 1). Since the beginning of 2020, approximately 17,000 more such papers have been archived.

Also, since the beginning of 2020, more than 6,000 manuscripts on coronaviruses have been uploaded to bioRxiv and medRxiv, two leading preprint servers. Recognizing the need to ensure open access to knowledge, the World Health Organization (WHO) created a preprint repository, COVID-19 Open, which had ~50 papers as of July 1, 2020. If papers deposited to COVID-19 Open
cannot find suitable publication elsewhere, they will be archived by the WHO and become citable as working papers.\(^{16}\)

While it is encouraging that the global research community has turned its attention to the COVID-19 crisis, and diverse opinions are important, not all papers are of equal merit or should be used to guide hypothesis design or public policy.

Findability, however, does not ensure quality. The publish-or-perish culture of academia is responsible for at least some of the failure of effective data sharing, even though the requirements of funding agencies promote data deposition in public repositories.\(^{17}\) Clinical and research data are often siloed, requiring further efforts to join them.\(^{18}\) Predictably, early COVID-19 papers have been retracted even from very high-profile journals,\(^{19}\) and reports of systemic quality issues are emerging.\(^{20}\) The Retraction Watch blog has created a special list for retracted COVID-19 publications.\(^{21}\)

**Discussion**

The science-policy interface fosters a virtuous cycle where each field improves the other, wherein access to clear and rigorous science is responsive enough to inform policy yet stringent enough to advocate for the needs of the scientist. That interface must be made more robust for research on outbreaks, whether measles in the US, Ebola in the Democratic Republic of the Congo, or a novel swine flu in China.

The global scope and global threat of pandemics underscores the need for a transnational entity such as the WHO, which can convene experts and funders to improve the infrastructure of outbreak response. The WHO is currently the only global institution with the reach, trust, and power to bring together the parties needed to build these resources and act as their arbiter. Here we propose two discrete actions, informed by previous successes in the field, to improve the invisible infrastructure of outbreak response. These improvements will directly advance the Research Priorities (Objective 5) outlined in the 2019 WHO Coordinated Global Research Roadmap R&D Blueprint.\(^{22}\)

Research would benefit from centralized repositories of data, both for the scale that they can achieve and for the inherent like-against-like comparisons that can be made with harmonized data. At present, the data used for display on each national dashboard are incomparable and their methodologies opaque. Closely linking the case data that underpins dashboards to pathogen sequence data could facilitate
computational experiments and the ability to develop and test hypotheses. However, the files behind these dashboards are incompatible with each other. The WHO global dashboard is also incompatible with country dashboards, and does not disaggregate the data by sex or age, important risk factors. There is a lack of transparency in the methods used for collection and aggregation, which would help transition these dashboards to research and policy resources.

For most researchers, these possibilities are presently limited by a lack of access to the relevant datasets, which is unproductive and undemocratic, and which furthermore stifles innovation and discovery. One model for expanding access to fine-grained case report data (the sort usually sequestered in medical records) is the Medical Information Mart for Intensive Care (MIMIC-III) database at the Massachusetts Institute of Technology. Another is the European Union’s COVID-19 data portal, which links virus and host genome, protein, expression, and biochemistry data, but leaves out clinical and case data. Adopting a system of credit sharing similar to GISAID’s would ensure that regardless of resource setting, data depositors are incentivized and recognized.

The centralization of outbreak data would have positive effects for policy- and decision-makers as they address local and global problems. With access to comparable global data, they can make better-informed decisions. However, they also need access to curated summaries of the literature written on the data, which will provide context and clues to which studies are superlative and should be used as a basis for action. Compendia of this sort will inform best practices for international consideration on policy for measures to flatten curves or open economies.

The removal of paywalls to promote access is not a new concept. There is a strong history of advocacy for broad access to knowledge, whether in recent commitments to open access in public health emergencies or the Findable Accessible Interoperable Reusable (FAIR) Data Principles from 2016, the Budapest, Bethesda, and Berlin statements on open access publishing and knowledge from the early years of the millennium, or the creation of the preprint arXiv in 1991. However, access without guidance can lead to chaos, and potentially a source of misinformation in the public sphere.

During the pandemic, these previously inward-facing resources have experienced increased interest and scrutiny from journalists and the general public. Creating a platform where data produced in the laboratory, field, and clinic face in to researchers and with summary, dashboard-like information facing out to decisionmakers will improve the quality and utility of those data across fields. Sharing credit more equally between depositors and users or analysts will
incentivize participation. Curating compendia for non-scientists will help bring the fields together; when they are paired with rapid peer review, the gap between new ideas and consumption by non-scientists across all fields is narrowed. Taken together, these data repositories and resources for high-quality summaries of impactful papers will help stimulate effective exchange at the science-policy interface.

**Recommendations**

Pandemics, as global problems, require harmonization across local contexts. Through better exchange of information at the science-policy interface, their human and economic price can be mitigated.

Broadly, we advocate that the WHO, under the auspices of Research Priority Objective 5,\(^3\) should take a lead role in fostering systems that aggregate case report data for outbreaks, modeling them after GISAID’s. We further advocate for the creation of metareview panels for papers related to the disease that would assemble and summarize strong research papers and preprints.

First, we propose a central repository for data from outbreaks, what that repository should contain, how to manage and perpetuate it, and possible limitations (Evidence & Science). Second, we explore the possibility for metareview and the creation of compendia of high-quality papers and summaries of those papers designed for a broad audience (Policy).

**A. Evidence & Science: Joint repository for case report and pathogen sequence data**

*Present assessment:* National-level dashboards cataloguing the spread of COVID are dynamic. In our initial survey (April 23, 2020), dashboards presented limited data, but as of July 1, the data and visualizations were generally more mature and of higher quality.

Publicly available data on COVID-19 require harmonization across sex, age, location, and disease severity. These data are of paramount importance in design of clinical trials and dosage levels for vaccines and therapeutics,\(^32\) the formation of basic hypotheses,\(^33\) and policy elaboration and implementation.\(^34\) Without a standard minimum amount of information included and a unified repository for these data, researchers cannot assess the biological and demographic factors that
affect health outcomes. The data are (largely) findable, but they are not accessible, interoperable, or reusable.

**Infrastructure need:** There is no single place to go for robust data on cases and the pathogen sequence. The resources that do exist are largely for expert scientists, but the need exists to span disciplines. This lack of infrastructure hinders our understanding of and our ability to respond to outbreaks. A repository, or set of repositories, with case reports and pathogen genome sequence data with an outward-facing dashboard designed for use by non-experts, and controlled access to the underlying data for researchers is needed.

**Informed execution:** The WHO Global Research Roadmap for COVID-19 provides an excellent, comprehensive of the research needs for this outbreak with an eye towards generalizable methodology, but little space is devoted to the mechanisms of data sharing.35

Rather than call on the WHO to build these platforms themselves, we suggest that it convenes stakeholders, who will themselves build consensus on the necessary data to be deposited into any repository and find funders to ensure stability of the resource, as is the case with GISAID. Stakeholders’ openness to ideas in less formal encounters, specifically at World Economic Forum meetings, were instrumental to building GISAID – this light-touch approach should be replicated and built upon by the WHO to build out similar systems for all outbreaks.

The stakeholders’ group should look to the genomics community and the NCBI Database of Genotypes and Phenotypes (dbGaP), a repository for personally sensitive or identifiable data.36 While often not user-friendly for submitters or accessors, it has helped to advance the state of the controlled access data sharing in genomics where some aspects of the data are individualized and identifiable. There, scientists eligible for an eRA Commons account may submit a project description to a Data Access Committee to access the data. eRA Commons is the United States National Institutes of Health investigator portal. This certification system is used to ensure a high bar of access for dbGaP and like resources. Similar controlled-access systems maintained by the European Bioinformatics Institute (EBI) and Chinese National Genomic Data Center also serve this purpose. Results of the parties convened could be reported with the same mechanism as the Global Research Roadmap and decision-makers could begin to form the public-private partnerships needed to address the agreed-upon gaps, needs, and challenges.

**Data deposition:** Support for centralized data sharing is growing and has a long history in genomics38 where it has improved the quality and scope of research. In our proposal for outbreak data sharing, it is reasonable to expect state agency data
depositions to come first, with researchers incentivized to deposit data thereafter, either at the time of publication or as a stipulation of funding.

Long-term support: This repository will require commitment to digital and human infrastructure that will take time to build, but which will have numerous long-term benefits and improve scientific study at all levels. GISAID, born of a public-private partnership, has been supported by numerous public entities. In 2010, the German Federal Republic took over housing the resource. The US, European Union, and China all support their federal data repositories, which are open for global use. A model that leverages governments, philanthropy and, potentially, commercial sector users will be key for the long-term sustainability of future endeavors.

Limitations of centralized data repositories

Privacy. The need to maintain essential liberties while enacting short-term, explicitly temporary restrictions proportionate with the risk to society is essential.\(^\text{39}\) It has been argued that COVID-19 provides an ideal case to test exceptions to existing data protection frameworks, including GDPR (General Data Protection Regulation).\(^\text{40}\) To this end, multiple tiers of access to a central repository could serve the greatest number of people while also modeling responsible privacy by design solutions, where proactively and transparently through all phases of engineering, measures are in place to protect the privacy and sensitivity of data is the default.\(^\text{41}\)

Hypothesis formation and evidence-based policy in the current pandemic rely on the accurate reporting of sex, age, and location data, which impact disease severity and control. To be useful in real time, one tier of data on these key aspects of the disease must face the public, while another tier of data remains private and available to researchers while reflecting local privacy laws. One dilemma is that the US HIPAA standard,\(^\text{42}\) which considers precise location and age after year 89 as personally identifying information. This information is critical to understanding this disease, yet reporting those data widely is irreconcilable with the current framework in the US. Furthermore, the heterogenous rule of law and standards of privacy across all nations will require rigorous debate over how to balance individual and collective needs. Where combined mobile device-based geospatial location data and contact tracing fits into this schema is unresolved, and warrants study in the bioethics and legal communities.\(^\text{43}\)

Equity. Health care systems are functionally different\(^\text{44}\) in high-, middle-, and low-income countries - we have seen this firsthand in our work in genetic medicine and infectious disease outbreaks in Central Africa.\(^\text{45}\) The use of and access to clinic and hospital settings, where data on epidemics are collected, varies with resources but also cultural norms. Infrastructure must be operable in low-resource settings...
with incentives for local participation to ensure that data from low-resource environments are accurately collected.

Interoperable data collection may also not be a high priority for staff focused on patient care in clinics. Our experience in clinical team science\textsuperscript{46} shows us that the more comprehensive the form, the less all fields are used. International data interoperability has another fundamental dilemma: harmonize data in English, or leave data in language with concepts useful to its generators, but more challenging to use by its aggregators? Here, we err towards equitability of language use, trusting that, with constant dialogue between the WHO and its regional and country offices, data will be human- or machine-processable and thus, interoperable.

**Accuracy.** Data collection can be hindered by agencies with overlapping mandates. For example, the Italian agency responsible for disaster mitigation documents different aspects of the pandemic than the Ministry of Health.\textsuperscript{47} This redundancy likely improves robustness but hampers data tracking for observers. Furthermore, faultlessly incorrect or intentionally misleading data are as detrimental as missing or incompatible data. This belies the need for *post hoc* data cleanup.

Our experience in coordinating large, multi-sited consortia (in English, French, and Spanish) has shown the need to strike a balance between easy-to-report, easy-to-collect, and easy-to-use data.\textsuperscript{48,49} Furthermore, our experience designing prescriptive templates to report disease-specific, standardized metrics, which ensure high-fidelity responses but require extensive training and practice, inform our recommendation to strike a balance between more and less open-ended responses in reporting cases.\textsuperscript{50} Creating new formats for data reporting is inherently a community activity, but remember that a camel is a horse designed by committee. Stakeholders must keep accuracy and utility at the forefront of their discussions. Once again, the WHO should provide leadership and guidance in constructing these much-needed data formats.

**B. Policy: Transparency at the science-policy interface through expert curation**

*Present assessment:* The viral growth of papers begs for comparison to be drawn to the outbreak. We would not suggest flattening this curve, as research and analysis of an outbreak that touches every corner of the earth are important. However, for scientists and policymakers alike, there is a need to sort the meritorious from the mere musings. Rather than more avenues for preprint publications, there is a need for curated libraries of science.
For maximum utility across fields, the glut of papers and preprints cannot simply be archived. High-quality items must be preserved and promoted with their systematic high-level expert summaries in a central hub. Those with preliminary, exploratory, or statistically underpowered findings must be handled carefully, and excluded where appropriate. Summary and archival of critical papers will serve policymakers and non-experts by bringing more arcane research to light. Rapid and centralized curation will help to stem the spread of misinformation and misunderstanding.

**Infrastructure need:** An endeavor of this ilk would require mobilizing wide-ranging expertise to assess the relevance of publications or preprints. One such strategy has been devised by a team at Johns Hopkins University who are offering metareview through their 2019 *Novel Coronavirus Research Compendium* (NCRC), which aims to provide “accurate, relevant information for global public health action by clinicians, public health practitioners, and policy makers.”\(^{51}\) Drawing upon PubMed and the preprint servers medRxiv, bioRxiv, and the Social Science Research Archive, as well as a suggestion box, they search for papers that best advance the state of COVID-19 knowledge to highlight and summarize.

On June 29, 2020, an initiative from MIT and UC Berkeley with a diverse editorial board of global experts launched a new journal, *Rapid Reviews: COVID-19,* “to combat misinformation in COVID-19.”\(^{52}\) “Promising” manuscripts identified in preprint archives by UC Berkeley’s COVIDScholar artificial intelligence tools will be subjected to expert review in a matter of days, and published with accompanying commentary. This model has the potential to revolutionize peer review as we know it and may become a model for all fields of research.

Such initiatives must be encouraged and globally funded, to ensure their independence and representativeness of high-quality research from across the world. For increased preparedness for the next pandemic, vetted compendia of the available literature should also be created for other viruses (e.g. Ebola, Zika), and other successive outbreaks. These metareviews should be housed in a central hub with easy access and promoted to policymakers.

**Leverage expertise:** Peer review relies on volunteer time from specialists in the field of the manuscript under review. Reviewers are contacted by editors on a case by case basis. For metareview and compendia, a more permanent group of volunteers or paid staff would be beneficial, activated as needed during outbreaks. These reviewers could be tapped and vetted by permanent staff, and offered compensation for their time, to ensure that even early- or mid-career academics could contribute. If reviewers had conflicts of interest, they would recuse themselves as usual.
Paradigm shift: Like peer review, curating and summarizing literature for a WHO-mediated compendium must become part of the job of the expert. This laborious, but significant, change at the science-policy interface will help increase the utility of open access knowledge. This work needs to be recognized by university promotion and tenure committees to ensure that experts tapped to do so can spend the necessary time and resources on this activity.

Conclusions

It is one thing to contemplate the needs of the world, and another entirely to know the world as it is - especially in a pandemic. COVID-19 has exposed the strengths and weaknesses of the many national and global systems of data sharing. When the doubling time of new papers is as fast as that of new cases of a virulent infection, it is not enough to have open access – guides and way finders are needed. As such, we propose solutions for better scientific data collection and dissemination. However, data availability only tells part of the story: resources must be effectively mobilized in order for policymakers to design the best responses. Many governments responded to the pandemic threat with protectionist action: by April 1, 2020, 68 jurisdictions had adopted export controls on everything from personal protective equipment to testing reagents and soap. We must work together to craft policy based in evidence, rather than protectionist fear if we are to win against COVID-19. While we have proposed solutions for data sharing in outbreaks and pandemics, a similar reflective analysis must be made by trade and policy experts. Efforts for data sharing and global policy design must be examined to improve flow of information and resources at the science-policy interface (Figure 2).
Figure 2. Graphic representation of the science-policy interface. Top and bottom text denotes flow of information as it cycles the scientific and diplomatic processes. In green, data inform the creation of new knowledge, which generates new data and improves old data. In blue, cooperation and transparency in decision-making combined with a barrier-free environment improve conditions to make better policy. These colors meet at the science-policy interface where facets of science and policy come together to inform each other, supporting an environment for evidence-based policy. In the COVID-19 pandemic, the flow of information across this interface broke down. In this manuscript, we describe structural improvements to the points illustrated by equilibrium arrows A and B. Formalized structures at these points will improve science and policy, while providing opportunity to ameliorate information flow at the science-policy interface. We call on the World Health Organization to act as convener to ensure effective, broad, and stakeholder-focused resources are created.

At equilibrium arrows A, we propose the construction of a tiered-access data repository to promote the flow of information to scientists and policy decisionmakers. Tiered-access will provide scientists with granular, identifiable data while also giving policymakers important high-level data. Some part of these data should be made open-access to promote transparency, while adhering to privacy by design.

At equilibrium arrows B, we propose the creation of metareview panels responsible for both the assessment of published literature and preprints, as well as the summary of those meritorious studies for non-science audiences.
About the authors

Jonathan LoTempio, BS, is a PhD candidate in the George Washington University Genomics and Bioinformatics program. His dissertation research is conducted at Children’s National Hospital in Eric Vilain’s laboratory. He previously served as a Scientific Program Analyst at the National Human Genome Research Institute and as a member of the Fast-Track Action Committee on Mapping Microbiomes of the United States Office of Science and Technology Policy.

D’Andre Spencer, MPH, is an epidemiologist who specializes in global health threats specifically in underserved communities. He is a staff scientist at the Center for Genetic Medicine Research at Children’s National Hospital where his work focuses on capacity building in health and genetic research in the Democratic Republic of the Congo and developing database infrastructure for multi-center translational research.

Rebecca Yarvitz is an international trade analyst in Washington, DC. She holds a BA in International Relations from the Elliott School of International Affairs of the George Washington University.

Arthur Delot Vilain is an independent high-school student with a focus on history, environmental sciences, and civil rights and social justice issues.

Eric Vilain, M.D., Ph.D., investigates questions at the intersection of science and society. He is the Director of the Center for Genetic Medicine Research at Children’s National Hospital, the Chair of the Department of Genomics and Precision Medicine at George Washington University, and the director of the CNRS International Research Laboratory “Epigenetics, Data, Politics”. He is an advisor to the International Olympic Committee Medical Commission and Past President of the International Academy of Sex Research.

Emmanuèle Délot, PhD, is a researcher and educator in developmental genetics and genomics. She is a professor at the George Washington University, and on the faculty at Children’s National Hospital. She serves as the founding executive coordinator for the Disorders/Differences of Sex Development Translational Research Network.

Acknowledgements

We would like to acknowledge and express gratitude for the thoughtful comments and suggestions from the external reviewers and the Science & Diplomacy editorial board.

This work was supported by the A. James Clark Distinguished Professor of Molecular Medicine Endowment to Professor Eric Vilain, MD, PhD, A. James Clark Professor, Children’s National Hospital.
We Can Do Better: Lessons Learned on Data Sharing in COVID-19... Jonathan LoTempio et al

Endnotes

2. For Iran, reports from the Islamic Republic News Agency were used in lieu of information from a public health agency.
11. Elbe and Buckland-Merrett, “Data, disease and diplomacy.”
16. Ibid.


29. “Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities,” https://openaccess.mpg.de/Berlin-Declaration


42. Mailman et al., “The NCBI dbGaP database.”


Annex

COVID-19 case data aggregated by national public health agencies as of April 23, 2020

<table>
<thead>
<tr>
<th>Country</th>
<th>Rank</th>
<th>Public Health Agency</th>
<th>Test</th>
<th>Contact</th>
<th>Exposure</th>
<th>Source</th>
<th>Location</th>
<th>Age</th>
<th>Sex</th>
<th>Severity</th>
<th>Death</th>
<th>Recovery</th>
<th>Reinfaction</th>
<th>Test Report</th>
<th>Interactive</th>
<th>CSV</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>1</td>
<td>Centers for Disease Control and Prevention</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Spain</td>
<td>2</td>
<td>Ministry of Health</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Italy</td>
<td>3</td>
<td>Higher Institute of Health</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>France</td>
<td>4</td>
<td>Public Health France</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Germany</td>
<td>5</td>
<td>Robert Koch Institute</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>6</td>
<td>Public Health England</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>Yes</td>
<td>Yes</td>
<td>ICU</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Turkey</td>
<td>7</td>
<td>TUBITAK</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Iran</td>
<td>8</td>
<td>News Agency</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>China</td>
<td>9</td>
<td>National Health</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Russia</td>
<td>10</td>
<td>Ministry of Health</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Brazil</td>
<td>11</td>
<td>Ministry of Health</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Singapore</td>
<td>12</td>
<td>National Health</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Canada</td>
<td>13</td>
<td>Public Health Agency of Canada</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Netherlands</td>
<td>14</td>
<td>National Institute for Public Health and the Environment</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>Hospitalization</td>
<td>ICU</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
<tr>
<td>Guatelmala</td>
<td>15</td>
<td>Federal Office of Public Health</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Only</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Web and PDF</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Superscript Country Link
c United States https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm?s_cid=mm6912e2_w
d United States https://www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm
e Spain https://covid19.isciii.es
g Italy https://www.epicentro.iss.it/en/coronavirus/bollettino/Infografica_15aprile%20ENG.pdf
i France https://www.data.gouv.fr/fr/datasets/donnees-des-urgence-hospitalieres-et-de-sos-medecins-relatifs-a-lepidemie-de-covid-19/
j France https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatifs-a-lepidemie-de-covid-19/
k France https://geodes.santepubliquefrance.fr/#c=indicator&view=map2
l Germany https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Fallzahlen.html
m Germany https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4
These are countries with the highest overall case counts on April 23, 2020 as collated by the Johns Hopkins University Center for System Science and Engineering. Each national public health website was queried on April 1, April 13, April 17, and April 23. As all websites reported the number of cases, a column was not created to document this metric in the table. In column 9, Severity, ICU stands for intensive care unit, the most common delineation of severe or critical cases. The Italian health ministry used a scale from asymptomatic to critical instead of hospitalization. Citations within the table document the URL accessed to find the information.
In column 9, Severity, ICU stands for intensive care unit, the most common delineation of severe or critical cases. The Italian health ministry used a scale from asymptomatic - critical instead of hospitalization. Citations within the table document the URL accessed to find the information. Note: all hyperlinks on both tables were verified and active at the time of submission, but may have expired since.