Data curation during a pandemic and lessons learned from COVID-19

Moritz U. G. Kraemer[1,*], Samuel V. Scarpino[2,3,*], Vukosi Marivate[4], Bernardo Gutierrez[1,5], Bo Xu[1,6], Graham Lee[7], Jared B. Hawkins[8,9], Caitlin Rivers[10], David M. Pigott[11], Rebecca Katz[12,*] & John S. Brownstein[8,9,*]

1. Department of Zoology, University of Oxford, Oxford, UK
2. Network Science Institute, Northeastern University, Boston, MA, USA
3. Santa Fe Institute, Santa Fe, NM, USA
4. Department of Computer Science, University of Pretoria, Pretoria, South Africa
5. School of Biological and Environmental Sciences, Universidad San Francisco de Quito, Quito, Ecuador
6. Department of Earth System Science, Tsinghua University, Beijing, China
7. Research Software Engineering Group, University of Oxford, Oxford, UK
8. Boston Children's Hospital, Boston, MA, USA
9. Harvard Medical School, Boston, MA, USA
10. Johns Hopkins Center for Health Security, Baltimore, MD, USA
11. Institute for Health Metrics and Evaluation, Department of Health Metrics Sciences, University of Washington, Seattle, WA, USA
12. Georgetown University, Washington, DC, USA

*Correspondence to -mail: moritz.kraemer@zoo.ox.ac.uk; s.scarpino@northeastern.edu; rebecca.katz@georgetown.edu; john.brownstein@childrens.harvard.edu

Detailed, accurate data related to a disease outbreak enable informed public health decision making. Given the variety of data types available across different regions, global data curation and standardization efforts are essential to guarantee rapid data integration and dissemination in times of a pandemic.

A wide range of data are critical to characterizing disease outbreaks and informing public health responses[1]. Pathogen genomic data have become essential to identify the causative agent of an infection, and they can also help track mutations and investigate transmission networks and the geographic spread of an infectious disease[2]. Clinical data are useful to understand disease severity, develop clinical case definition, evaluate pharmaceutical interventions and monitor disease outcomes[3]. Serological data are important to characterize immunity, antibody responses and how they may relate to clinical outcomes[4]. Additionally, epidemiological data ranging from aggregated case counts[5] to detailed contact-tracing data have been used extensively to characterize basic parameters such as the reproduction number, key time distributions (onset of symptoms to hospitalization among others), and heterogeneity in transmission[6]. Metadata associated with individual epidemiological cases can be of great importance to understand early disease dynamics and critical transitions from imported infections to those locally acquired[7]. Further, disease data that contain demographic information have been used extensively to understand population level attack rates[8]. All of these data inform response actions designed to mitigate the consequences of the disease event.

Whereas many countries and jurisdictions collect detailed data during outbreaks, they may not be shared openly due to various ethical, legal and privacy issues, political regulations and concerns, and/or computational limitations[9]. Computational frameworks for rapid ingestion of these data are not widely available either. In addition, there are no standardized data formats that facilitate the open reporting of such information while ensuring its compliance with regulations around data privacy (primarily de-anonymization). This makes international comparisons of large, detailed outbreak data difficult and prevents inferences from such data to be effective in response to disease outbreaks. As a consequence, there is a missed opportunity in having a single platform where all of the data, irrespective of type and region, can be easily and quickly shared among the scientific community, which can greatly accelerate research related to a disease outbreak.

During the COVID-19 pandemic, the Global.health initiative was established to create a global infrastructure for consolidating, standardizing and sharing individual-level epidemiological data across different geographic regions. Nevertheless, challenges related to data ingestion and curation still persist, and addressing them is of crucial importance to enable rapid data analysis of open data during future outbreaks.

Data standardization

There are substantial challenges when ingesting epidemiological data from across the world mainly due to the diversity in formats that data is collected, summarized and disseminated. Therefore, a critical step to facilitate data sharing among different countries and communities is to define a standard format that is common across all the different health reporting systems. Because data can come from different sources, this standard format must accommodate multiple data streams, such as those from official government sources, news outlets and social media. In addition, this format must be pathogen agnostic, meaning that it should be extensible to multiple infectious diseases in order to facilitate adoption in future outbreaks.

In the Global.health initiative, a standardized format for epidemiological data[10] was agreed on by different regional working groups after consulting a number of public health agencies, academic research groups, and health policy experts. The main goal during these interviews was to identify a format that could accommodate most use cases, enable rapid decision making and reduce time to clean data. This included mapping data to a common geographic reference frame. International collaborations, as well as sharing of testing protocols and local expertise, were essential to define a standard: these collaborations must continue in the future to make sure that we have usable standards for the community. Finally, the standardized format was developed with extensibility in mind, although this needs to be periodically revised to make sure that the resources in place can include pathogen-specific information from other infectious diseases.

There are still some challenges to be solved. For example, symptom ontologies vary substantially across regions and limited amount of information about country specific triaging protocols were available. In addition, observational data are likely to suffer from substantial biases in how data are collected. Better ways to characterize these different biases in the standardized format will be particularly important in future outbreaks.

Data ingestion workflow

Entering data by hand is certainly unfeasible beyond ~50,000 cases and as an outbreak grows exponentially. Therefore, automatic workflows for data ingestion are essential to allow rapid ingestion of data. They need combined expertise of data scientists and epidemiologist working alongside each other.

A series of programmatic approaches were developed in the context of Global.health to allow ingestion of data from structured and unstructured sources. Importantly, these approaches can be applied to datasets ranging from PDF documents to standard application programming interfaces (APIs), and can be performed manually and in more automated ways. Designing more integrated workflows that enable rapid communication and exchange between public health agencies could greatly enhance the understanding of the data-collection process and its associated limitations.

Data quality

Data quality may vary substantially across different data streams, especially early during an outbreak, when data are often less structured. Also, biases may only become apparent after they have been collected and analyzed. Therefore, while researchers and decision makers need to perform their own assessments of the feasibility of specific data to support their study findings, it is imperative to have a data validation process in place for the ingestion.

For future outbreaks, it will be important to implement validation workflows in which human curators look at data streams at a daily level. In addition, the data must indicate whether an entry has been verified by a curator, to better guide users. Scalability here can be a challenge: as the data grows, manual validation can become unfeasible. Nevertheless, a decentralized model, where volunteers and team members across different regions perform data validation, can help alleviate this challenge, as shown by the COVID Tracking Project. More automated workflows for validation can also be implemented to assist human curators (for example, anomaly detection algorithms).

Even though we believe that a decentralized model will be the most effective in building a trusted community, we do acknowledge the need for stable funding and a well-organized and transparent informatics and administrative center, especially during the start of a pandemic, when the nature of data is uncoordinated and often disparate. Flexible approaches will be needed to accommodate them, especially as every outbreak is different. Furthermore, and because all infectious disease outbreaks have the risk of becoming global, a globally comprehensive database will be helpful in guiding coordinated responses.

Data integration

Key data that are collected during a public health investigation for infectious disease management are not limited to demographic and clinical information. Other important data may be serological data, pathogen genomic data, or non-epidemiological, spatial data that help characterize drivers of transmission (these are usually socio-economic, demographic and environmental). The underlying process of data generation (at which resolution, which study population and so on) varies across these different types and integrating them sensibly is a major opportunity for improving infectious disease research.

Remarkable innovations have been made to make such data available at a global scale and at fine spatial resolution[11]. To accommodate research and inferences of disease dynamics that pair epidemiological data with spatial data, the common geographic reference frame defined in the standardization process will help to merge data across data types. However, computational platforms that allow researchers to easily pair spatial data (for example, population density) with epidemiological data from a given geographic region will enable researchers with little experience in spatial data processing to perform complex and integrated data analyses.

Outlook

The collection, integration and dissemination of timely high-resolution global epidemic data has the opportunity to augment public health response and inform policy in real-time. Trust will be one of the key components enabling rapid data sharing and the misuse of data has been detrimental to data sharing and disincentivized open collaborations. We advocate for better principles around the terms of use of data and general principles of data sharing that must include guidelines that prevent data use to reinforce existing biases or discrimination against specific populations based on gender, age or location. Furthermore, appropriate computational infrastructure will need to be developed so that the risk of re-identification is minimized and balanced against the potential impact on health of the wider population.

It is also crucial that all of the data and code are open source to enable rapid integration across multiple research groups and governments in the future. A truly open platform can assist users in overcoming existing geographical, organizational and societal barriers to information access, and enable great public health empowerment and democratization.

Acknowledgements

We thank all members of the Open COVID-19 Data Working Group (https://github.com/beoutbreakprepared/nCoV2019). The Global.health initiative was supported by a philanthropic gift from Google.org and a grant from the Rockefeller Foundation.

Contributions

Conception and design: M.U.G.K., S.V.S., J.S.B., D.M.P., R.K., B.G. and C.R. M.U.G.K. wrote the first draft of the manuscript. V.M., B.X., G.L. and J.H. edited the manuscript. All authors contributed to revisions and approved the manuscript.

Ethics declarations

Competing interests

The authors declare no competing interests.

References

1. Lipsitch, M., Swerdlow, D. L. & Finelli, L. *N. Engl. J. Med.* **382**, 1194–1196 (2020).
2. Grubaugh, N. D. et al. *Nat. Microbiol.* **4**, 10–19 (2019).
3. Salje, H. et al. *Science* **3517**, 208–211 (2020).
4. Kucharski, A. J. & Nilles, E. J. *Lancet Infect. Dis.* **20**, 1351–1352 (2020).
5. Dong, E., Du, H. & Gardner, L. *Lancet Infect. Dis.* **20**, 533–534 (2020).
6. Lewnard, J. A. et al. *BMJ* **369**, m1923 (2020).
7. Kraemer, M. U. G. et al. *Science* **368**, 493–497 (2020).
8. Rodriguez-Barraquer, I., Salje, H. & Cummings, D. A. *eLife* **8**, e45474 (2019).
9. Xu, B. & Kraemer, M. U. G. *Lancet Infect. Dis.* **20**, 534 (2020).
10. Xu, B. et al. *Sci. Data* **7**, 106 (2020).
11. Kraemer, M. U. G. et al. *Trends Parasitol.* **32**, 19–29 (2016).