Perspective

Following Data as it Crosses Borders During the COVID-19 Pandemic

Joseph M. Plasek, PhD[1#], Chunlei Tang, PhD[1#*], Yangyong Zhu, PhD[2], Yajun Huang, PhD[3], David W. Bates, MD, MSc[1]

[1]Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02120, USA.

[2]School of Computer Science, Fudan University, Shanghai 201203, CHN.

[3]School of Economics, Fudan University, Shanghai 200433, CHN.

[*]Correspondence: ctang5@partners.org

[#]JMP and CT contributed equally.

Keywords: Health information interoperability; Global health; Medical economics; Data science

Words: 2,027; 148 in Abstract

Abstract: Data changes the game in terms of how we respond to pandemics. Global data on disease trajectories and the effectiveness and economic impact of different social distancing measures are essential to facilitate effective local responses to pandemics. COVID-19 data flowing across geographic borders are extremely useful to public health professionals for many purposes such as accelerating the pharmaceutical development pipeline, and for making vital decisions about intensive care unit rooms, where to build temporary hospitals, or where to boost supplies of personal protection equipment, ventilators, or diagnostic tests. Sharing data enables quicker dissemination and validation of pharmaceutical innovations, as well as improved knowledge of what prevention and mitigation measures work. Even if physical borders around the globe are closed, it is crucial that data continues to transparently flow across borders to enable a data economy to thrive which will promote global public health through global cooperation and solidarity.

Tracing the origins of new diseases through their growth into global pandemics, such as the 2019 RNA virus strain from the Coronaviridae family known as COVID-19, necessitates following the flow of relevant data. Two weeks after the first COVID-19 hospitalization, virologists conducted metagenomic RNA sequencing on a patient and published its molecular blueprint (a dizzying string of more than 34,000 letters) about a month later[1-2]. News reports and other biosurveillance related data pointed to a cluster of pneumonia cases that an AI-driven algorithm called BlueDot identified as being an outbreak on December 31st, 2019, a week before global public health officials notified the public[3]. Outbreaks, such as on the Diamond Princess cruise ship, provided valuable information about how the disease is spread and its incubation period[4-5]. Electronic health record systems have augmented their self-reported travel screening questionnaires to help identify patients who have recently visited areas where community spread is present[6].

Transportation data have been used to simulate the spread of a disease and estimate the effect of local and intercontinental travel restrictions[7]. Air, sea, and land transport networks continue to expand in reach, speed of travel, and volume of passengers carried, providing a vector for infectious disease spread. Simulations suggested cancelling the Spring Festival in China – a period known for crowded buses, trains, planes, and ferries culminating in an estimated 3 billion trips. Prescriptive analytics on outbreak data through algorithms or models can simulate possible outcomes and help answer: "what should we do" when the outbreak constitutes a public health emergency of local or international concern. Decision-making about travel advisories and quarantines is done locally, and each locale has its own level of preparedness for an outbreak. Some areas have used innovative approaches; for example, Taiwan integrated its health insurance database with biometric entry and exit data to generate real-time alerts based on travel history and clinical symptoms to aid in case identification and has used this data to decide whom to quarantine and track at the border[8]. The global health security index (GHSI) encompasses disease prevention, detection, reporting, and response capabilities for each country. Countries with a higher GHSI like Singapore can identify undetected cases through increased epidemiological surveillance and contact-tracing, leading to improved accuracy regarding disease prevalence[9].

There are many potential international data sources for disease surveillance systems[10] to utilize. For example, ProMED Mail, a program of the International Society for Infectious Diseases, is useful for monitoring emerging diseases. Aggregators of local media reports and news feeds, such as DXY, Google News, Baidu News, SOS Info, and Moreover are useful in identifying new cases early on. Monitoring social media feeds like Facebook and Twitter as well as trends in Google search terms can be useful to have an idea about what is present in a community or what people are worried about. Tracking animal, agriculture, and environmental health data for potential sources of human disease is possible via the

Wildlife Data Integration Network (WDIN), the Food and Agriculture Organization of the United Nations (FAO), and the World Organization for Animal Health (OIE).

A common way to disseminate data about infections like COVID-19 is through data visualizations and simulated disease models. These data products enable the public, policy makers, and scientists to quickly understand the global spread of COVID-19 at the population level, enabling forecasting at the local level. HealthMap[11-12] provides an accurate, continuously updated, and usable visualization tracking the global spread of COVID-19 over time. The John Hopkins dashboard tracks cases in different geographic areas across the globe[13]. The Institute for Health Metrics and Evaluation (IHME) has developed a real-time data visualization and forecasting tool based on geo-coded epidemiological information that includes (when available): symptoms, key dates (date of onset, admission, and confirmation), and travel history[14].

These examples of data and data product flow across geographic borders are extremely useful to public health professionals for many purposes such as accelerating the pharmaceutical development pipeline, for triaging clinician resources to a locale, and for making decisions about intensive care unit rooms, where to build temporary hospitals (e.g., Boston Hope Medical Center[15]), or where to boost supplies of personal protection equipment, ventilators, or diagnostic tests[16]. Providing data analytical tools for organizations that cannot share data or have limited analytical resources can also be helpful to help with virus response, better-coordinated care, reporting, and organizational operations. The health sectors in advanced economies can help developing countries via cross-border data product sharing, as they did with the Congo in the 2018 Ebola outbreak. This included early screening (e.g., outbreak detection), continuing disease surveillance, advice regarding travel advisories, and ex situ medical treatment (e.g., medical tourism) and helped result in improved quality of care, and reductions in cost.

Global virtual hackathons focused on COVID-19 such as the Observational Health Data Sciences and Informatics (OHDSI[17]) international communities' study-a-thon (March 26-29, 2020) and the Massachusetts Institute of Technology hackathon (April 3-5, 2020) have also spurred the development of cross-border clinical research studies and cross-border entrepreneurship, respectively. The output of these efforts are open source ideas and tools to solve a variety of problems arising from the COVID-19 pandemic. The OHDSI efforts are focusing on a global baseline characterization of COVID-19 patients as well as the safety of Hydroxychloroquine for COVID-19 treatment, drawing upon collaborators and data spread across the globe.

The flow of COVID-19 data across borders also has economic implications. In the field of biomedical informatics, we sometimes ignore the economic effects that the data and data products we create and consume may have on the global economy, but it is worth examining them in the context of a global pandemic. Certainly, COVID-19 has had a devastating effect on the global economy, and that

could affect public health in a variety of ways[18]. From a health data economy perspective, the inflows and outflows of data and information across geopolitical boundaries has the potential to generate enormous economic value in a digitally connected global healthcare economy[19-20]. The capital value of global COVID-19 data can be maximized when analyzed using descriptive, predictive, or prescriptive analytics for the purposes of clinical research, public health purposes, and pharmaceutical development. These cross-border data flows have the potential to be a driver of global economic growth, though altruism has largely dictated the free flow of data and ideas in the current crisis. COVID-19 data enables significant new opportunities for innovation and disruption within the health data economy, especially for emerging infectious diseases[20], and telemedicine. Governance of data and data product sharing can take the form of the OHDSI network with the free flow of data products for a collective research gain, a commercial data sharing model such as between Google and HCA Healthcare[21], or a self-governance model like DataBox[22] where profit sharing from the data transfer can be realized by the data owners.

The goal of flattening the curve is to reduce the reproduction ratio. The reproduction ratio is how many people that a person in one disease episode passes the disease along to the next group (e.g., if the reproduction ratio is four, then that infected patient transmitted COVID-19 to four more people[23]). However, the number of reported cases may not be a very useful indicator unless you know something about how the COVID-19 testing is being conducted and how the data are being gathered[23]. When there are major differences between COVID-19 testing strategies, as there has been in this pandemic, it is difficult to make direct comparisons accurately as the testing strategies can skew case counts[23]. Accurate estimation of the reproduction ratio depends on having comprehensive, diverse, and heterogeneous data sets to overcome the limitations of individual localized data sources. For COVID-19, countries that conducted comparatively high numbers of tests had lower mortality rates even though they reported high case counts that alarmed the public in the short run[23]. Tracking the viral mutations of COVID-19 cases in New York suggests that most cases were traced to travelers returning from Europe, not Asia as originally expected[24]. Missing this hidden spread due to insufficient testing and screenings at the borders meant that the suspension of air travel and mandatory quarantines for travelers from Europe occurred too late.

Global data on disease trajectories and the effectiveness and economic impact of different social distancing measures are essential to facilitate effective local responses to pandemics. Policymakers have used these data to inform their decisions regarding travel bans, quarantines, and economic stimulus. To facilitate the dissemination of knowledge regarding COVID-19 during the outbreak, publishers are prioritizing review of and offering free, open access to relevant research findings[25]. Sharing COVID-19 data freely and globally boosts the data economy, enabling quicker dissemination and validation of pharmaceutical innovations, as well as improving knowledge of what prevention and mitigation

measures work. Even if physical borders around the globe are closed, it is crucial that data related to COVID-19 continue to transparently flow across borders to enable a data economy to thrive which will promote global public health through global cooperation and solidarity.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Competing Interests Statement: The authors have no competing interests to declare.

Contributorship Statement: Tang C, Zhu Y, Huang Y, and Bates DW built on and extended the initial idea. Tang C drafted this manuscript. All authors provided substantial contribution to paper edits. Plasek JM, especially, filled up this manuscript with great content to increase its size. All the authors are accountable for the integrity of this work.

Acknowledgements: The authors would like to thank Sheng Wang, PhD for valuable comments and suggestions on early drafts. The content is solely the responsibility of the authors.

References

1 Wu F, Zhao S, Yu B, et al. A new coronavirus associated with human respiratory disease in China. Nature. February 3, 2020. DOI: 10.1038/S41586-020-2008-3.

2 GenBank: MN908947.3. Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome. National Center for Biotechnology Information. January 23, 2020. Retrieved from https://www.ncbi.nlm.nih.gov/nuccore/MN908947.

3 Niiler E. An AI epidemiologist sent the first warnings of the Wuhan virus. Wired.Com. January 25, 2020. Retrieved from https://www.wired.com/story/ai-epidemiologist-wuhan-public-health-warnings.

4 Apuzzo M, Rich M. Yaffe-Bellany D. Failures on the Diamond Princess Shadow Another Cruise Ship Outbreak. NYTime. Retrieved from https://www.nytimes.com/2020/03/08/world/asia/coronavirus-cruise-ship.html.

5 Zhu N, Zhang D, Wang W, et al. A novel cornavirus from patients with pneumonia in China, 2019. N Engl J Med. 2020;382:727-733. DOI: 10.1056/NEJMoa2001017.

6 Miliard M. Epic pushes out software update to help spot coronavirus. Healthcare IT News. January 24, 2020. Retrieved from https://www.healthcareitnews.com/news/epic-pushes-out-software-update-help-spot-coronavirus.

7 Chinazzi M, Davis JT, Ajelli M, et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. March 6, 2020:eaba9757. DOI: 10.1126/science.aba9757.

8 Wang CJ, Ng CY, Brook RH. Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. JAMA, March 3, 2020, DOI:10.1001/jama.2020.3151.

9 Niehus R, Salazar PMD, Taylor A. et al. Quantifying bias of COVID-19 prevalence and severity estimates in Wuhan. China that depend on reported cases in international travelers. medRxiv, 2020, 13:2020.

10 Mandl KD, Overhage JM, Wagner MM, et al. Implementing syndromic surveillance: a practical guide informed by the early experience. J Am Med Inform Assoc. 2004;11(2), 141-150.

11 HealthMap. Boston Children's hospital. Retrieved from https://www.healthmap.org/covid-19.

12 Kracmcr M. I'm a researcher who's helped change how we tackle pandemics like coronavirus forever – this is what we've learned. Independent.co.uk. March 17, 2020. Retrieved from https://www.independent.co.uk/voices/coronavirus-covid-19-pandemic-outbreak-data-research-cdc-who-a9406281.html.

13 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. Lancet Infect Dis. February 19, 2020. doi: 10.1016/S1473-3099(20)30120-1.

14 Xu B, Gutierrez B, Mekaru S, et al. Epidemiological data from the COVID-19 outbreak, real-time case information. Sci. Data. March 24, 2020. doi:10.1038/s41597-020-0448-0.

15 Erickson JI. "Boston Hope" Medical Center Opens at Boston Covention and Exhibition Center. March 10, 2020. MassGeneral.org. Retrieved from https://www.massgeneral.org/news/coronavirus/boston-hope-medical-center-opens.

16 Pagel C, Utley M, Ray S. Covid-19: How to triage effectively in a pandemic. March 9, 2020. The BMJ opionion. Retrieved from https://blogs.bmj.com/bmj/2020/03/09/covid-19-triage-in-a-pandemic-is-even-thornier-than-you-might-think.

17 COVID-19 updates page. Observational health data sciences and informatics. Retrieved from https://www.ohdsi.org/covid-19-updates.

18 Frazee G. How the coronavirus' economic toll could also affect public health. March 30, 2020. Retrieved from https://www.pbs.org/newshour/economy/making-sense/how-the-coronavirus-economic-toll-could-also-affect-public-health.

19 Dobbs R., Manyika, J., Woetzel J. Digital globalization: The new era of global flows. McKinsey Global Institute. March 20, 2015. Retrieved from https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Globalization/Global%20flows%20in%20a%20digital%20age/Global_flows_in_a_digital_age_Full_report%20March_2015.ashx.

20 Tang C, Plasek JM, Bates DW. Rethinking Data Sharing at the Dawn of a Health Data Economy: A Viewpoint. JMIR, 2018;20(11): e11519. https://doi.org/10.2196/11519.

21 Kent J. HCA, Google Cloud Launch COVID-19 Data Sharing Platform. HealthITAnalytics.com. April 09, 2020. Retrieved from https://healthitanalytics.com/news/hca-google-cloud-launch-covid-19-data-sharing-platform.

22 Zhu Y, Xiong Y, Liao Z, et al. Self-governing openness of data. Big Data Research. 2018;4(2):3-14, in Chinese.

23 Silver N. Coronavirus case counts are meaningless. FiveTirthEight.com. April 4, 2020. Retrieved from https://fivethirtyeight.com/features/coronavirus-case-counts-are-meaningless/?fbclid=IwAR1gpC1Zblt_rPRfBkMUZKBVNNKSTrIS3WcS_P6gwE1uGvCp98CspXRCzTE.

24 Zimmer C. Most New York Coronavirus Cases Came From Europe, Genomes Show. NYTime. April 08, 2020. Retrieved from https://www.nytimes.com/2020/04/08/science/new-york-coronavirus-cases-europe-genomes.html.

25 Calling all coronavirus researchers: keep sharing, stay open. Nature. 2020;578:7. DOI: 10.1038/d41586-020-00307-x.