A need for open public data standards and sharing in light of COVID-19

The disjointed public health response to the COVID-19 pandemic has demonstrated one clear truth: the value of timely, publicly available data. The John Hopkins University (JHU) Center for Systems Science and Engineering's COVID-19 dashboard[1] exists to provide this information. What grew from a modest effort to track a novel cause of pneumonia in China quickly became a mainstay symbol of the pandemic, receiving over 1 billion hits per day within weeks of its creation, primarily driven by the general public seeking information on the emerging health crisis. Critically, the data supporting the visualisation were provided in a publicly accessible repository and eagerly adopted by policy makers and the research community for purposes of modelling and planning, as evidenced by the more than 1200 citations in the first 4 months of its publication. 6 months into the pandemic, the JHU COVID-19 dashboard still stands as the authoritative source of global COVID-19 epidemiological data.

Similar commendable efforts to facilitate public understanding of COVID-19 have since been introduced by various academic, industry, and public health entities. These costly and disparate efforts around the world were necessary to fill the gap left by the lack of an established infrastructure for real-time reporting and open data sharing during an ongoing public health crisis.

Although existing systems were in place to achieve such objectives,[2–4] they were not empowered or equipped to fully meet the public's expectation for timely open data at an actionable level of spatial resolution. Moving forward, it is imperative that a standardised reporting system for systematically collecting, visualising, and sharing high-quality data on emerging infectious and notifiable diseases in real-time is established. The data should be made available at a spatial and temporal scale that is granular enough to prove useful for planning and modelling purposes. Additionally, a critical component of the proposed system is the democratisation of data; all collected information (observing necessary privacy standards) should be made publicly available immediately upon release, in machine-readable formats, and based on open data standards.

Developing such a system will be subject to numerous challenges. Based on our experience of COVID-19 data collection and reporting, specific issues may include: (1) the ambiguity of parameter definitions and variable reporting frequencies; open, standardised reporting protocols are necessary for the operators of such a distributed system (eg, local, state, and federal public health agencies in the USA) to ensure consistency in data structures, parameters, and metadata requirements, necessary for seamless aggregation and processing of the data. (2) Discrepancies in reporting across sources; the system architecture should be hierarchical in nature, wherein data collection is sourced at the local level (municipalities or counties), then aggregated up to state or province, country or region, and globally, further minimising redundant data collection efforts between agencies with overlapping jurisdictions (ie, local and state governments). (3) Inconsistencies in reporting medium; the system architecture should rely on common standards and distribution mechanisms. (4) Adaptability to changes in reporting over time; the system will require flexibility in the data structure and effective communication channels. (5) Privacy concerns; all personally identifiable data should be anonymised and aggregated, and privacy should not be a reason for delaying or withholding the public release of data. The system should be hosted by an apolitical organisation, sufficiently resourced and empowered, and supported by a global community of public health experts.

The development and implementation of such a system is not a simple task. However, it is feasible and far more economical and efficient than the duplicative efforts required by third parties to build such a platform in real time amid a global health crisis, as was the case with COVID-19. Critically, without such a system in place, our ability to respond effectively to the inevitable future pandemics will be limited, and the cost will be human lives.

*Lauren Gardner, Jeremy Ratcliff, Ensheng Dong, Aaron Katz
l.gardner@jhu.edu

Department of Civil and Systems Engineering and Center for Systems Science and Engineering, Johns Hopkins University, Baltimore, MD 21218, USA (LG, ED); Applied Physics Laboratory, Johns Hopkins University, Laurel, MD, USA (JR, AK); and Nuffield Department of Medicine, University of Oxford, Oxford, UK (JR)

1   Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; **20:** 533–34.
2   WHO. Coronavirus disease (COVID-2019) situation reports. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed July 22, 2020).
3   ECDC. Situation updates on COVID-19. https://www.ecdc.europa.eu/en/covid-19/situation-updates (accessed July 22, 2020).
4   US CDC. National syndromic surveillance program (NSSP). https://www.cdc.gov/nssp/overview.html (accessed July 22, 2020).