Improving epidemic surveillance and response: big data is dead, long live big data

Epidemics pose a growing threat. Our cities are increasingly densely populated, we are more connected than ever before, and in recent years we have witnessed successive waves of new (severe acute respiratory syndrome [SARS], Zika virus, Ebola virus, and now coronavirus disease 2019 [COVID-19]) and old (influenza) infectious disease threats causing global pandemics. Urgent investment in surveillance systems and global partnerships are needed to prepare for the pandemics that will continue to emerge in the coming decades. There has been discussion of the promise of integrating sophisticated epidemiological models and new big data streams—for example, from mobile phones, satellites, or social media—at various stages of the public health response, particularly in the context of epidemic forecasting and decision making.[1] These new data streams provide important, real-time information about travel patterns that spread disease and spatial shifts in populations at risk, which until recently have been very difficult to quantify on timescales relevant to a fast-moving epidemic. With growing mobility and increasing global connectivity, this information will be key to planning surveillance and containment strategies.

In theory, with appropriate data sharing protocols in place, it should be possible to produce useful, up-to-date epidemic forecasts informed by these data streams. For this process to be effective, individuals from different institutions, including academia, industry, non-governmental organisations (NGOs), and governments, need to be in frequent communication. Key privacy concerns must be addressed for the routine use of new data streams, in particular the most appropriate way to robustly aggregate these data streams to ensure the anonymity of individuals. But even if privacy is addressed, there are additional structural challenges to the translation of new approaches in a decision-making context. Here, I focus on three of these challenges as they pertain to creating useful epidemic forecasts during an outbreak.

The first challenge is that incentives across the analytical pipeline are misaligned.[2] Academics are largely incentivised to write scientific articles and to fund their work through individually led grants. These activities are not conducive to the rapid response to a crisis (although many academics do respond), or to sustained engagement and training of corporate and government teams. Companies are incentivised by profit, and are rightly beholden to national regulatory frameworks and the public with respect to the data they collect. Ministries of health have complex relationships with both the companies that have access to personal data and with the public. They face many competing health priorities and complex political choices when it comes to open sharing of epidemiological data. Disease control programmes are often hampered by limited capacity and high turnover of personnel, and health workers on the ground during an epidemic have their hands full responding to the immediate crisis and might not be trained or incentivised to report epidemiological data accurately. Taken together, these incentive structures create multiple barriers to rapid data generation and the development of streamlined epidemic forecasting systems that use these new data types.

The second challenge for implementing real-time epidemic forecasting is the gap between technological or methodological innovation, which often occurs in academic settings in high-income countries, and implementation in field settings, frequently done by NGOs or governments in low-income and middle-income countries. Many funders have adopted a financing model intended to spur innovation through short-term pilot projects that place greater emphasis on the novelty of a technology or method than on the validation of its impact, but this approach exacerbates this separation. Pilot funding also fails to acknowledge the long timelines required to engage with health systems effectively and to measure health impact rigorously, instead promoting one-off bilateral collaborative projects that do not scale up and are not sustained after the lifetime of the project. Much more investment is needed in the validation, implementation, and scaling up of innovations in close intellectual partnership with stakeholders who are responsible for delivering them,[3] rather than continuing to fund the proliferation of solutions divorced from the problems themselves. In a recent editorial, Seye Abimbola noted

that "while the gulf between discovery and delivery exists in other fields, what makes global health peculiar is that discoveries and the decisions on whether or how to deliver them are typically made at a distance, removed from the realities of their targets or intended beneficiaries."[4]

The third challenge is methodological: epidemic forecasting is inherently uncertain. There is sometimes an underlying assumption in the big data and artificial intelligence (AI) narrative that complex simulation models and mobile phone data or statistical covariates can bypass the need for the collection of basic epidemiological information. However, for emerging outbreaks—with COVID-19 highlighting this point—we often lack accurate data about case counts and biological processes driving an epidemic, let alone the behavioural responses of people affected, making it challenging to swiftly adapt or interpret very complex models on the spatiotemporal scales relevant for decision making. Arguably, the most useful frameworks will tend to be simple,[5,6] both because of the need for flexible models that yield rapid answers given the large uncertainty surrounding epidemiological data during an emergency, and because simple models are more easily interpreted and communicated.[2] Clear communication of both the value and limitations of model outputs is a prerequisite to their useful deployment, but is often absent. Since policy makers generally do not have in-depth modelling expertise, lack of clear communication risks two negative outcomes: believe models without scepticism and decisions will be misinformed, or dismiss modelling out of hand and fail to use the evidence we have to contain outbreaks as effectively as possible.

Decisions must be made quickly during epidemics based on patchy and uncertain data, and models can be powerful tools to help guide them. Despite the challenges above, ongoing advances in computational power, methods, and new data streams offer genuine hope for better surveillance and useful forecasting systems. New data sources at our disposal include not only the passively observed big data streams from mobile phones but also detailed environmental data and local sensor information from distributed devices, internet search information, pathogen genomic data that can be generated rapidly during an outbreak to inform the response,[7] and crowd-sourced approaches to monitoring rapidly evolving emergencies.[8] Data

sharing platforms and standardised aggregation approaches that protect the privacy of personal data are being developed,[9] and increasing internet connectivity allows for rapid data transfer and communication between geographically disparate teams of responders. Methodologically, powerful ensemble modelling approaches are being developed that combine multiple forecasts to minimise uncertainty.[5,10] We have seen an unprecedented, collaborative approach unfolding in response to the COVID-19 outbreak between academic groups: for example, using Twitter and other platforms to share, analyse, and openly discuss the implications of new data as they come out. (Ironically, managing the disinformation that also proliferates on social media during emergencies will probably become one of the most important issues for epidemic containment moving forward.)

These innovations will remain dislocated and impractical until the challenges above are addressed. Encouragingly, all three issues could be improved by moving much of the focus of funding and expertise to the populations most vulnerable to epidemics. The unpredictable nature of epidemics and the increasingly technical components of these approaches means that flexible, distributed teams of people who span the analytical and operational aspects of the outbreak response will be needed if we are to use new big data approaches to complement and clarify the small data—epidemiological, geographical, and social—that are essential for the development of meaningful forecasts. Regional or local teams with analytical training and ongoing relationships with government and industry partners should lead to flexible modelling approaches that leverage large curated datasets and international analytical expertise, building on collaborations developed when there are no emerging or ongoing outbreaks.[5] This approach could also help to alleviate some of the political issues associated with data sharing, even when crucial information cannot be shared publicly. For this concept to work, substantial long-term investment is needed to train and support individuals, particularly in low-income and middle-income countries, to fill what is currently a major gap in the analytical pipeline. Indeed, if we are to confront future epidemics with the best tools at our disposal, a new kind of interdisciplinary training—distinct from traditional Master of Public Health (MPH)

and graduate programmes—is needed, as well as the development of non-traditional public health careers, to foster individuals who can direct flexible forecasting efforts, engage with governments and industry partners, validate approaches, and create channels for communication.

I declare no competing interests.

Caroline Buckee
cbuckee@hsph.harvard.edu

Center for Communicable Disease Dynamics, Harvard TH Chan School of Public Health, Boston, MA 02115, USA

1   George DB, Taylor W, Shaman J, et al. Technology to advance infectious disease forecasting for outbreak management. *Nat Commun* 2019; **10:** 3932.

2   Kahn R, Mahmud AS, Schroeder A, et al. Rapid forecasting of cholera risk in Mozambique: translational challenges and opportunities. *Prehosp Disaster Med* 2019; **34:** 557–62.

3   Gates B. The next epidemic—lessons from Ebola. *N Engl J Med* 2015; **372:** 1381–84.

4   Abimbola S. On the meaning of global health and the role of global health journals. *Int Health* 2018; **10:** 63–65.

5   Viboud C, Sun K, Gaffey R, et al. The RAPIDD ebola forecasting challenge: synthesis and lessons learnt. *Epidemics* 2018; **22:** 13–21.

6   May RM. Uses and abuses of mathematics in biology. *Science* 2004; **303:** 790–93.

7   Armstrong GL, MacCannell DR, Taylor J, et al. Pathogen genomics in public health. *N Engl J Med* 2019; **381:** 2569–80.

8   Schiavo R, May Leung M, Brown M. Communicating risk and promoting disease mitigation measures in epidemics and emerging disease settings. *Pathog Glob Health* 2014; **108:** 76–94.

9   de Montjoye YA, Gambs S, Blondel V, et al. On the privacy-conscientious use of mobile phone data. *Sci Data* 2018; **5:** 180286.

10   Ray EL, Reich NG. Prediction of infectious disease epidemics via weighted density ensembles. *PLoS Comput Biol* 2018; **14:** e1005910.