Productive disruption: opportunities and challenges for innovation in infectious disease surveillance

Caroline O. Buckee,1 Maria I E Cardenas,2 June Corpuz,3 Arpita Ghosh,4 Farhana Haque,5 Jahirul Karim,6,7 Ayesha S. Mahmud,1 Richard J Maude,1,7,8 Keityl Mensah,1 Nkengafac Villyen Motaze,10 Maria Nabaggala,11 Charlotte Jessica Eland Metcalf,9 Seda Aurélien Mioramalala,12 Frank Mubiru,11 Corey M. Peak,1 Santanu Pramanik,4 Jean Marius Rakotondramanga,13 Eric Remera,14 Ipsita Sinha,7,8 Siv Sovannaroth,15 Andrew J Tatem,16 Win Zaw7

BACKGROUND

Infectious diseases place an unacceptable and disproportionate social and economic burden on low-income countries. National disease control programmes have the difficult task of allocating limited budgets for interventions across regions of their countries, based on often disparate datasets of varying quality from a range of sources including clinics, hospitals, village health workers, the private sector and non-governmental organisations (NGOs). Every stage of the data collection and analysis pipeline for surveillance systems may be affected by a lack of capacity as well as by biases and misaligned incentives for reporting and managing data. Addressing these issues will be essential for effective reduction in the burden of endemic infectious diseases globally as well as to preparing for emerging epidemic threats.

Meanwhile, academic researchers—often in high-income settings—are developing increasingly sophisticated methods to collect and analyse data to improve spatial estimates of disease burden using new Big Data sources, mobile-Health or m-Health approaches or mechanistic and statistical modelling techniques. While these advances leap ahead, however, many remain most useful for estimating global disease distribution,1 rather than for national control programme prioritisation. Translating these new techniques to inform policy in endemic settings remains challenging. The pronounced disconnect between health systems and academia may limit the utility of new approaches. The high burden of work placed on healthcare workers in low-income settings further limits their scope and time available for engagement with methodological developments.

Despite ongoing challenges to implementation, however, there are promising analytical approaches that can leverage even patchy and low-quality data and diverse new data streams that can be productively harnessed to strengthen strategies for resource allocation when integrated with existing surveillance systems. We detail the data and analysis challenges faced by national disease control programmes, outline possible solutions offered by analytical approaches and new data-streams and conclude by outlining barriers to implementation.

Summary box

► New innovations that could transform infectious disease surveillance and control, including the use of Big Data, mobile health approaches and cutting edge quantitative methods, offer hope for disrupting traditional health systems and improving health worldwide.
► Much has been made of their potential, but very few have been translated successfully into policy or scaled up to a population level.
► We argue that there is currently a lack of integration of new approaches, making them unsustainable or unrealistic for most national control programmes and that the gulf between academia and policy makers remains a major barrier to their implementation.
► We propose that these innovations must be designed with direct input from national control programmes and embedded within already existing health systems.
CHALLENGES ASSOCIATED WITH INFECTIOUS DISEASE SURVEILLANCE SYSTEMS

Generally, epidemiological data about patients are reported by healthcare practitioners via passive surveillance systems to a central database, which is used to determine trends over time in and map the geographic distribution of burden of disease in different regions as well as the extent and efficacy of interventions (active surveillance via sentinel sites may also inform these efforts). These regional data in turn serve as an important basis for resource allocation decisions. Figure 1 illustrates the flow of data and potential hurdles faced by national control programmes and the ways in which new approaches may be used in parallel to traditional systems.

Data quality is often perceived as a major barrier to using passive surveillance data to guide resource prioritisation. Data may be delayed or simply missing from core settings and, even where available, is frequently underreported, potentially in spatially and temporally variable ways. For example, a problematic reporting pattern that emerges repeatedly is an apparent increase in disease incidence that is actually caused by increased surveillance efforts and/or diagnostic capacity. Over-reporting is also a potential hazard when local regions are financially incentivised to exaggerate their needs. Further, since many local health centres lack diagnostic capabilities, much of the large-scale data rely on syndromic surveillance (influenza-like-illness, diarrhoea, fever) with low specificity. Additionally, estimates of the catchment population of health facility or district may be flawed, since data may be based on a single census from up to a decade ago, thus potentially both out of date and a poor reflection of seasonal fluctuations in population numbers. Collection of more detailed, highly curated data at sentinel sites seems promising, but may add little to national decision-making because such high-quality sites are likely to be limited in their spatial scale. All these issues mean that where case numbers exist, policy makers may have little confidence in them.

LEVERAGING FRAGILE DATA USING STATISTICAL AND MATHEMATICAL PRINCIPLES

The first question is what can be done with existing infrastructure and data. With the right statistical or theoretical tools, even low-quality data can potentially be leveraged to help inform strategic deployment of control efforts. Thoughtful deployment of interpolation or geostatistical tools can be used to create smooth maps of burden or intervention efforts across space, also...
allowing extrapolation to unmeasured contexts.\textsuperscript{1,4,5} Autocorrelation models are also powerful tools, building on surveillance data to guide predictions about outbreaks of dengue, for example.\textsuperscript{6} Moving from statistical to mechanistic approaches, even if incidence reporting is erratic, dynamical signatures of the infectious process might still be detectable if additional data on features of cases are available, such as age, geographic location and gender. Age is a powerful covariate for infectious disease dynamics, as age of infection is linked to the magnitude of transmission. High-transmission pathogens often have a low average age of infection, as they move quickly through immunologically naïve populations.\textsuperscript{7} Conversely, implementation of control efforts is likely to increase the average age of infection,\textsuperscript{8} so intervention efficacy may be measured using shifting age structure of cases.

Where data are consistent through time, but not space, basic principles from infectious disease dynamics open the way to estimating characteristics of pathogens. For example, the growth rate of an epidemic can be extracted from incidence, allowing estimates of the net reproduction number, or \( R_0 \), which captures the degree to which an outbreak is expected to grow (\( R_0 > 1 \)) or shrink (\( R_0 < 1 \)) (eg, deployed during the recent Ebola outbreak despite variation in reporting rates).\textsuperscript{9} Allocation of resources towards ‘source’ populations, where \( R_0 > 1 \) then becomes possible—although maps of the locations and densities of rural populations (ie, denominator challenges) are also necessary. Infectious disease models can allow characteristics of the surveillance system, such as the magnitude of under-reporting, to be estimated where the susceptible population can be inferred (eg, via susceptible reconstruction).\textsuperscript{10} Where only syndromic surveillance is available, it may be possible to correct for background rates of focal syndromes to pull out the dynamics associated with a particular infection. This strategy has been used for influenza\textsuperscript{11} (influenza-like-illness data are frequently available but data on influenza are rare), enabling investigation of signatures of climate effects on the burden of infection,\textsuperscript{12} for example, which has potential to contribute to planning efforts. Alternatively, simulation tools based on known epidemiological parameters of particular pathogens can establish the degree to which the data are reliable and forecast outbreaks and emergence events and/or the impact of interventions like vaccination (eg, roll-out of cholera vaccination to contain an epidemic).\textsuperscript{13}

\textbf{COMPLEMENTING EXISTING EPIDEMIOLOGICAL INFORMATION WITH NEW DATA SOURCES}

While these analytical strategies can compensate for the limitations of different data quality issues, a range of promising new data streams are also available. Rapid technological advances make these increasingly affordable, offering additional or new data layers to include in epidemiological analyses.\textsuperscript{11,14,15} We focus on three new data types.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Optimal use of new approaches depends on epidemiological context. Different phases of epidemiological containment and control lend themselves to different analytical approaches and data sources. Here, we have highlighted the spatial dimensions of this issue, with emergence and elimination phases exhibiting high spatial heterogeneity. In these cases, pronounced heterogeneity produces signals in data that can be leveraged to model the spread of infection between populations. For endemic infections where prevalence is distributed throughout the country and controlling disease burden is the primary purpose of interventions, the use of age profiles of exposure and other analytical approaches may be used to enhance or make use of patchy or poor quality data.}
\end{figure}
that are tractable in the context of control programme capacity: geospatial data, passively collected mobile phone records and pathogen genomic data. Note that we do not discuss the many mHealth approaches to actively engaging with populations directly, for surveillance or for interventions like health education, but these also provide data that do not rely on traditional surveillance systems. In general, different needs dominate at different phases across an epidemiological spectrum from emergence to elimination (figure 2), which will determine which data are useful and how they should be analysed.

Geospatial data are increasingly of high resolution and encompasses settlements and transport networks, indices of vegetation coverage, land use, land surface temperature and wind speed. These data can be combined with other geospatially referenced data such as meteorological data from weather stations, population densities or road networks to generate comprehensive estimates of environmental variables or indicators like remoteness or urbanicity relevant to communicable disease transmission. This can then be combined with point pattern data on vector presence via machine learning algorithms to determine features such as the likely range and local transmission intensity of vector borne infections like Zika virus.

Mobile phone data—routinely collected by operators and providing information about the location and movements of subscribers in real time—offer tremendous promise for control programmes to measure disease spread, if appropriately anonymised and aggregated to protect subscriber privacy. Indeed, the development of appropriate anonymisation and aggregation protocols remains an important priority for academics and public health practitioners and will require carefully balancing the ethical risks of reidentifying individuals with the benefits of predicting disease spread and identifying targets for intervention. Integration of this information into risk mapping by control programmes offers particular promise where disease incidence is heterogeneous in time and space and mobility drives both changing risk. This can then be combined with point pattern data on vector presence via machine learning algorithms to determine features such as the likely range and local transmission intensity of vector borne infections like Zika virus.

ENGAGING COMMUNITY HEALTH WORKERS AND RESEARCHERS WHO COLLECT AND ANALYSE DATA IS KEY

While our focus is on the potential of existing data despite perceived inadequacies and in use of novel data streams, it is important to note that country experience points to many low-cost adjustments to practice that could improve data availability. In multiple settings, there are important opportunities to share data across diseases, breaking down silos between disease surveillance and/or control programmes. This effort would be enhanced by electronic rather than paper-based reporting systems, allowing standardised data collection, aggregation and sharing. Perhaps most importantly, however, effectively engaging community health workers and others who collect and report patient data will be key to strengthening surveillance. For example, in Cambodia, malaria case reporting was improved when healthcare workers received feedback about the incidence of malaria in their area using colour-coded stickers; the effective use of mobile phones and tablets for reporting cases improved when health workers could use them for personal use and where internet coverage was restricted, the flexible reporting using mobile phones via SMS and later smartphones greatly improved timeliness of data collection.

It will be essential to sustain resources for engaging with health workers involved in generating and reporting surveillance data and not to impose technological advances at the expense of motivating and retaining qualified, reliable workers who form the foundation of an effective health system. Barriers to implementing approaches that use novel types of data or new analytical tools are rooted in human capacity (figure 1) at every level within control programmes, from local healthcare centres up to the ministry of health. While a considerable strength of novel data-streams such as satellite or mobile phone data are that they bypass many of the levels of reporting where data may be lost and incentives are misaligned, the insights they provide will be limited for setting public health agendas if they are not integrated
with traditional reporting systems. For academics, one of the most important contributions in this regard will be providing training, so that public health officers have the technical capacity to understand the benefits of, and oversee the implementation of, new methods. Sustainable integrating of these approaches will require sustained commitment from Ministries of Health, donor agencies and sustained.

**Author affiliations**

1. Department of Epidemiology, Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, Massachusetts, USA
2. Centro de Atención y Diagnóstico de Enfermedades Infecciosas, Universidad Industrial de Santander, Bucaramanga, Colombia
3. Epidemiology Bureau, Department of Health of the Philippines, Manila, Philippines
4. Public Health Foundation of India, Vasant Kunj, New Delhi, India
5. Programme for Emerging Infections (PEI), Infectious Diseases Division (IDD), ICDDR, B, Dhaka, Bangladesh
6. Disease Control Department, Directorate General of Health Services, Dhaka, Bangladesh
7. Mahidol-Oxford Tropical Medicine Research Unit, Faculty of Tropical Medicine, Mahidol University, Bangkok, Thailand
8. Centre for Tropical Medicine and Global Health, Nuffield Department of Medicine, University of Oxford, Oxford, UK
9. Dept of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ, USA
10. Division of Epidemiology and Biostatistics, Stellenbosch University, Stellenbosch, South Africa
11. Infectious Disease Institute, College of Health Sciences, Makerere University, Uganda
12. M&E Service, Data & Malaria Survey Division, National Malaria Control Program, Antananarivo, Madagascar
13. Epidemiology Unit, Institut Pasteur, Madagascar
14. Rwanda Biomedical Center, Kigali, Rwanda
15. Technical Bureau, National Malaria Control Program, Pnom Penh, Cambodia
16. WorldPop, Department of Geography and Environment, University of Southampton, Southampton, UK

**Contributors** All authors conceived, planned and contributed intellectually to this work. COB and CJEM wrote the manuscript, and all authors edited the manuscript.

**Funding** This work was the result of the Harvard-Wellcome Trust Workshop on Infectious Disease Modeling for Control Programs in 2017. It was supported by Wellcome Trust Fund 106866-Sustaining Health Award and by Award Number U54GM088558 from the National Institute of General Medical Sciences.

**Disclaimer** The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

**Competing interests** None declared.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data sharing statement** No additional data are available.

**Open Access** This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/

**REFERENCES**