Innovations in Disease Surveillance and Monitoring

13

Jessica Espey and Hayden Dahmm

Contents

J. Espey (✉) · H. Dahmm
UN Sustainable Development Solutions Network, London, UK
e-mail: Jessica.espey@unsdsn.org; hayden.dahmm@unsdsn.org

Abstract

While there have been worldwide improvements in public health over the past few decades, new viruses and other pandemics are expected to increase as we move towards high-density, urban living. But with only 49 countries reporting high-quality cause of death data to the WHO, the exact burden of infectious diseases is impossible to know. As made clear by the COVID-19 pandemic, there is a vital need for identifying and monitoring the spread of new infectious diseases. Novel sources of data offer the opportunity to expand the geographic scope and timeliness of surveillance activities. This chapter reviews six new health surveillance approaches, such as syndromic data collection, as well as three other novel data collection methods, such as movement tracking using telecommunications data, that have the potential to offer useful health insights. However, most of these methods are in their comparative infancy and need rigorous testing and investment before a pandemic strikes so that countries are not left scrambling to innovate during a crisis period. Particularly important is that countries aim to establish an inclusive national data infrastructure, which fosters third party partnerships and innovation, with epidemic intelligence as core objective of that system. The focus should be data-driven and evidence-based decision-making during crisis periods as well as day-to-day policy-making.

Keywords

Infectious diseases · Monitoring · Surveillance · Big data · Evidence-based public health

Introduction

There have been substantial declines in the death rates for a number of infectious diseases over the past decade. In 2017, 15% of deaths were attributed to communicable diseases compared with 25% in 1990 (IHME 2020a). Yet with the growing threat of antibiotic resistance, there has been a marked increase in the number of emerging diseases (WHO 2020). Furthermore, new viruses and other pandemics are expected as we move towards high-density, urban living (UNEP 2020). But with only 49 countries reporting high-quality cause of death data to the WHO, the true burden of infectious diseases is impossible to know exactly (IHME 2017). As made clear by the COVID-19 pandemic, there is a vital need for identifying and monitoring the spread of new infectious diseases. Novel sources of data offer the opportunity to expand the geographic scope and timeliness of surveillance activities.

Our knowledge about the global distribution of diseases is still limited (Hay et al. 2013). As of 2013, only 7 of 355 significant infectious diseases had been comprehensively mapped, and not knowing the geographic distribution of disease has real public health implications (Ibid.). Moreover, the Ebola outbreak, which killed over 11,300 in West Africa (Bell et al. 2016), underscored the fundamental weaknesses in

national and international systems for disease detection, monitoring and response (Woolhouse et al. 2015), made all the more apparent by the present coronavirus epidemic.

New data collection methods and data sources can improve the quality and quantity of information available for disease surveillance systems, if carefully managed, by providing:

- more timely data to inform policy and decision-making;
- more accurate spatial maps to support preparedness and planning;
- and more data with which to derive insights on disease spread patterns and extent (Hay et al. 2013).

Traditional Disease Surveillance Practices

International Health Regulations, established in 2005, define public health surveillance as "the systematic, ongoing collection, collation and analysis of data for public health purposes and the timely dissemination of public health information for assessment and public health response as necessary" (CDC 2006a). But the data collected by such systems is not exclusively health related, it may include other demographic, socioeconomic and clinical characteristics of the population under surveillance, data on key outcomes such as disease complications and mortality, and data on potentially mitigating or aggravating behaviors or co-morbid conditions referred to as risk factors (Soucie 2012).

National communicable disease surveillance systems around the world provide the majority of this critical data, although these systems have sometimes been developed unevenly, with surveillance conducted by a variety of agencies, or far from the point of action, or by academic institutions rather than by government (WHO 1999). The WHO has defined a set of core functions provided by a surveillance system: case detection; reporting; investigation and confirmation; analysis and interpretation; action; control and response; policy; and feedback (WHO 1999). Although the system as a whole covers a wide range of functions, data and information are fundamental to case identification, monitoring and evaluation.

To date, health information systems continue to rely on a range of manually collected datasets at both national and regional levels (Bansal et al. 2016). The three most common sources are clinical data, civil registration and vital statistics data, and administrative data from health facilities, discussed below. The collection of data from these sources is broadly guided by the following principles, which borrow from the Principles of Official Statistics (Soucie 2012):

- Surveillance must have a clearly defined objective, which will dictate what data is most relevant.
- Measurement standards are critical and definitions must be clear from the start.
- Standardization of the data collection is essential for comparing population groups, geographic areas, or trends over long periods of time.

- All data elements should be clearly defined so enumerators or clinicians are all collecting comparable information.
- The minimum amount of data should be collected to meet the stated surveillance objective.
- Large and complex data collection tools should be limited where possible, as they can substantially increase the burden of data collection which may adversely affect both the amount and quality of the data collected.
- Careful targeting of the population is essential, including a strong sampling strategy if the data are to be representative.
- Data should be gathered using an appropriate information gathering style (e.g., patient interview, clinical record review) such that the responses will most likely be valid and the data reliably reflect the true status of the condition under study.
- In all cases it is extremely important to apply ethical principles during the collection of data and to respect the privacy of the individuals under surveillance.
- Laws and regulations concerning the confidentiality of data collected are universally available and should be adhered to as a matter of standard practice.

Clinical Data

Disease outbreaks are normally detected by clinical investigations by healthcare workers (Woolhouse et al. 2015). For example, in a number of countries, "sentinel surveillance systems" have been established in health care sites such as hospitals, clinics, or care providers' offices to monitor key disease trends such as influenza or cancer. The main purpose these provider-based surveillance systems are "to obtain timely information on changes in the occurrence of a disease or condition that can inform preventive public health activities" (Soucie 2012). Clinical surveillance data was crucial to the identification of the West African Ebola epidemic from 2013 to 2016 (WHO Ebola Response Team et al. 2014).

Yet there are severe limitations to surveillance purely through clinical investigations; first it needs very strong government and institutional support, with standardized data collection techniques used by all clinicians (Soucie 2012). Second, there are medical practice challenges, such as mild symptoms not being detected in clinics or being misdiagnosed (Campos-Outcalt et al. 1991). Alternatively, clinics may be underutilized, as sick people choose not to visit a clinic because of cost or distance. For example, in parts of sub-Saharan Africa, such as Ethiopia, upwards of 18% of women do not have access to essential health services as they are more than 2 h away (as of July 2020) (World Pop 2020).

Furthermore, traditionally monitoring of disease outbreaks using clinical data would involve health officials taking reports of known disease occurrence, drawing on the literature. This information would be combined with expert knowledge about environmental factors, such as temperature and rain fall, to map and predict disease risk. This process can be highly labor intensive and has often been done manually (Hay et al. 2013).

CRVS Data

Civil registration and vital statistics data (CRVS) can be a useful source of public health data. A well-functioning system records all births and deaths, including cause of death, as well as other vital events such as marriages and divorces. It is defined by the UN as the "universal, continuous, permanent and compulsory recording of vital events provided through decree or regulation in accordance with the legal requirements of each country" (UNSD in WB & WHO 2014). Unlike other sources of vital statistics, such as censuses and household surveys, the data from CRVS systems enables "the production of statistics on population dynamics, health, and inequities in service delivery on a continuous basis, for the country as a whole and for local administrative subdivisions" (WB & WHO 2014) (Fig. 1).

In spite of its obvious benefits, many countries do not have comprehensive CRVS systems (more than 100 worldwide) which means that an estimated two-thirds of deaths are never recorded (WB & WHO 2014). As a result, their deaths are not included in vital statistics systems, inhibiting accurate information about the spread of diseases. Likewise, it is estimated that approximate 230 million children under five have not had their birth registered either, which means that even if their deaths are reported, there will be no record of their lives beforehand, including where they lived, what services they've accessed and their movements, which all have implications for monitoring disease spread (WB & WHO 2014).

Administrative Data

Finally, administrative data can also be a source of disease surveillance data. Encompassing a range of data collected by the government for non-statistical purposes, administrative data often includes healthcare, insurance, and education data. Administrative data, especially related to health care, has been used to track the prevalence and outcomes of diseases. Health administrative data usually entails hospital billing records, primary and secondary diagnoses, procedure codes, provider names, admission and discharge dates, and demographic information on the individual.

One example of a type of administrative data is a registry. In health these can be patient registries, medical or health ministry registries. While these tend not to include detailed clinical data, they can be a useful source of information for population wide, representative data on common diseases and conditions. Where registries and other administrative data are recorded electronically, are interoperable, and shared across government departments, they can be a particularly useful resource for monitoring national health and wellbeing.

Existing Limitations

As intimated above, these traditional approaches help us paint a broad picture of disease behavior but are subject to a number of limitations.
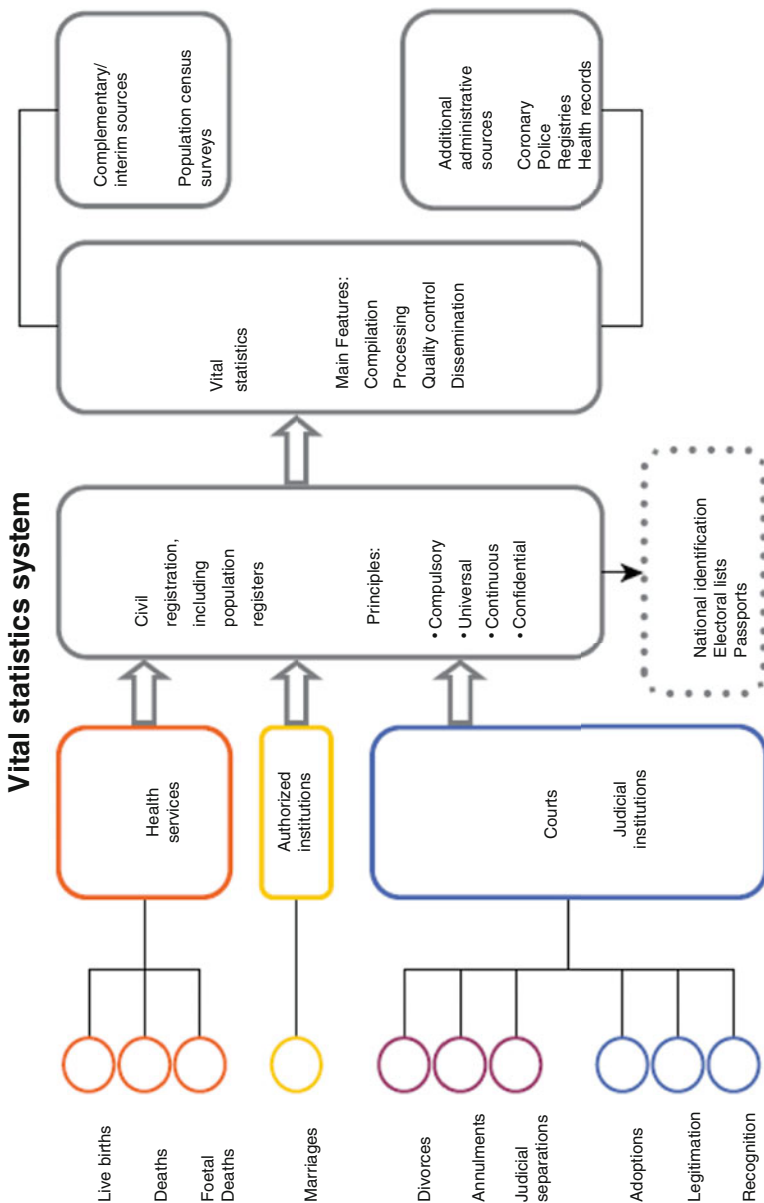
Fig. 1 A civil registration and vital statistics system. (Source: United Nations (2014) Principles and Recommendations for a Vital Statistics System, Revision 3, United Nations: New York)

The first is underfunding; existing systems are not funded sufficiently to fulfil their missions (Espey 2015). CRVS and administrative data sets are often incomplete, in part due to a lack of funding that inhibits the complete collection of this data and the maintenance of systems for accessing the data. According to the last recorded estimate, produced in 2014, the cost of scaling up CRVS systems to 73 priority countries was estimated to be US$3.82 billion over a 10-year period (2015–2024) and at the time of writing the financing gap was 52%, an estimated $1.98 billion (World Bank & WHO 2014). Recent estimates of the cost of monitoring COVID-19 in the USA suggest that surveilling that one disease across the U.S could cost as much as $3.6 billion in additional emergency funding to state and territorial health departments, suggesting huge additional shortfalls in most other countries worldwide (Watson et al. 2020).

A related issue is timeliness; many traditional health data systems are not updated with enough frequency to help with disease surveillance. In order for clinical, CRVS, and administrative data to be useful in disease surveillance, they must be able to provide data as close to real-time as possible. The time lag inherent in these traditional approaches also limits their utility in helping to rapidly identify and respond to disease outbreaks. Many of these methods, especially clinical data, are collected and coded manually, leading to a delay in their availability for analysis. In terms of disease mapping, this traditional approach produces maps that are spatially continuous but only represent one point in time (Hay et al. 2013).

Underreporting also presents limitations to traditional disease surveillance. Because much of the data for surveillance is based on passive reporting by healthcare providers, only a fraction of cases is ever reported. The Centers for Disease Control outlines two main reasons for underreporting – lack of knowledge about reporting requirements and negative attitudes towards reporting. If individuals are unaware of processes, if the reporting system is too burdensome or if there is distrust of the public health system, cases are less likely to be reported. Underreporting can delay treatment and lead to the spread of a disease (CDC 2012).

New Sources of Health Surveillance Data

The ongoing data revolution has the potential to strengthen surveillance systems by improving the timeliness of data and bringing new sources of information to bare. This ranges from data about patient symptoms to digital data from web searches and social media, as well as telecommunications data about population movement and satellite data that can describe relevant environmental changes. Aside from new technologies, building capacity and simply changing the procedures for data collection is just as important, as with efforts by the UN to improve autopsy reporting. New data sources, including big data, have the potential to greatly improve the timeliness and availability of epidemiological data, but these new sources are intended to strengthen and supplement traditional systems, not replace them (Bansal et al. 2016).

Strengthening Traditional Data with New Methods

The most urgent demand on CRVS data in the health sector is to provide reliable, standard data on cause of death (Ye et al. 2012). Even though mortality data is essential to documenting diseases, overworked health systems can struggle to deliver this data systematically, and many deaths happen outside of the hospital setting. The WHO has established methods for conducting verbal autopsies, which are 15-minute interviews with witnesses of deaths. This data can then make needed contributions to CRVS (WHO 2017). In particular, verbal autopsies can provide information about disease incidents like malaria. The INDEPTH project (International Network for the Demographic Evaluation of Populations and their Health) has demonstrated the benefits of this verbal autopsy method in Asia and Africa. In a study of 22 INDEPTH sites classifying nearly 100,000 deaths, researchers were able to track mortality rates and causes of depth and found the results to be comparable to other sources of mortality data. Although verbal autopsies are generally not as conclusive as clinical data, in the aggregate, they provide a useful and low-cost way of identifying trends in mortality in the absence of CVRS data (Streatfield et al. 2014). This shows that we don't necessarily need new technologies or new types of data. Alternative methods for collecting recognized data can be of great benefit.

> **Box 1: Verbal Autopsy Methods**
> Verbal autopsies are a method of determining causes of death and associated health concerns in deceased individuals where there is an incomplete vital registration system. Usually a trained interviewer will undertake a questionnaire with an individual familiar to the deceased, asking about signs, symptoms, possible risk factors (like smoking), and demographic characteristics. Once collected the questionnaire responses are customarily reviewed by a physician but this can be costly and time-consuming, as such groups like IHME and Johns Hopkins have been trialing alternative review processes such as "InterVA" (a program that incorporates commonly applied physician decision points by coding them into algorithms).
>
> IHME concludes that "a great deal of research has been conducted in the past several decades about VA, but some traditional methods of implementation and analysis can be costly, time-consuming, and potentially of varying quality. Verbal autopsies are now analyzed using a much wider array of cutting-edge techniques, some of which could be less expensive or yield higher quality results than those used traditionally," but these developments are a work in progress (IHME 2020b).

Point of Care Data

Point of care (POC) data are produced by rapid diagnostic tests linked to IT systems. Results are inputted electronically by providers while care is being administered to patients. It allows for quick and low-cost diagnosis while requiring limited

infrastructure (Kozel and Burnham-Marusich 2017). Improving detection times can be vital to disease control (Woolhouse et al. 2015) and POC data has the potential to do so in both high- and low-resource settings.

Because there was no vaccine available, the main way to control the 2014 Ebola outbreak in West Africa was through containment, which required rapid diagnosis in order to isolate patients. Real time testing for the Ebola virus was developed in the early 2000s and compared with traditional methods; it took hours rather than days to obtain results. However, at the start of the outbreak, these tests had not been approved for clinical settings. New tests had to be developed for use in decentralized health care facilities with minimal laboratory infrastructure, use of minimally invasive diagnostic samples, and simple procedures. When they were made available, these new tests were able to provide more rapid results than traditional testing, leading to faster isolation of patients (Broadhurst et al. 2016).

The shortage of POC tests for COVID-19 in the United States compared with countries like South Korea has shown the importance of being able to rapidly test and isolate patients. Early on in the outbreak in the United States, testing shortages meant that most of the people who were able to get tested were already very ill and that asymptomatic or mild cases were not being identified and isolated. Meanwhile, South Korea was testing tens of thousands of people per day with rapid results and the cases in South Korea have stayed low relative to the United States. These two countries' experiences with point of care testing demonstrate its effectiveness at rapidly responding to outbreaks (Kost 2020).

Sero-Surveillance

Sero-surveillance is a technique that is used to measure population level exposure (Woolhouse et al. 2015). Immunity and past exposure to diseases can be determined by testing for antibodies to a disease. Serological surveillance can detect antibodies even without symptoms or an active virus. It can provide a broad picture of population immunity and how a disease is spreading, as well as identifying immunization coverage gaps and predict outbreaks (Arnold et al. 2018). For existing epidemics, serological data shows which populations have been affected and which may still be susceptible while also providing inputs for models that can determine transmissibility and severity.

Box 2: Serological Studies on H1N1
During the H1N1 epidemic in 2009, a large number of serological studies were used to provide a picture of how the virus had spread throughout the world and determine infection rates. These studies were valuable for understanding the virus but the majority were not published until the virus was widespread. As a result, information about the disease's transmissibility and severity were not available early on when it could have been useful in preventing the spread of

> **Box 2:** (continued)
>
> the disease. Challenges involved in scaling up serological studies early on included the development of new procedures, procurement of funding, rapid deployment of training, and laboratory capacity. However, in the case of H1NH, once the serological studies ramped up, researchers were provided with information that allowed them to learn more about the virus and its spread. Lessons from serological testing during the H1N1 epidemic include adopting a common framework, methodology, and reporting system, standardizing laboratory serological procedures and planning for outbreak studies (Laurie et al. 2013).

Sero-surveillance has also been used during the COVID-19 epidemic to understand the spread of the disease and the infection rate within the population. The Centers for Disease Control plans to use this data to see how many infections have occurred at different points in time, at different locations, and among different populations. The results of this testing will help guide future control measures and healthcare needs by understanding the incidence of infection, how it spreads, and what populations are most vulnerable. It can also help identify risk factors for the disease such as age, location, and underlying health conditions. Given the shortage of testing in the United States early on in the pandemic, serology testing may also help to identify asymptomatic and mild cases that were not tested while they had the virus (CDC 2020).

Symptomatic Data

Symptomatic data is data about the health symptoms reported by patients, rather than formal diagnoses. Participants are asked on a weekly basis to report on their symptoms such as fevers, cough, body aches, diarrhea or vomiting. Often, this informal data is collected electronically and can be used to detect spikes in certain symptoms among a broader population. This preliminary information can help in tracking outbreaks (Pilot et al. 2011). This type of surveillance is most often used to detect influenza via online apps and websites and are based on volunteer participants. Flu Near You uses volunteers in the USA to monitor for influenza-like-illness in which the volunteers report on the prevalence of 10 symptoms every week. A website shows the distribution of symptoms throughout the country, allowing for early detection. However, in order to be most effective, systems like *Flu* For Now need a large and diverse sample size (Chunara et al. 2013).

> **Box 3: The Indian Integrated Disease Surveillance Project**
> In India, responsibility for disease surveillance is shared by central and local government authorities (Pilot et al. 2011). Responding to issues with

Box 3: (continued)

variability in consistency, the Indian Integrated Disease Surveillance Project was set up, with WHO support, in 2004 to collect village level data on a weekly basis about symptomatic reports. In the city of Pune, a symptomatic screening process was set up in response to H1N1, mapping the location of patients, and identifying patterns in where patients originated (Pilot et al. 2011).

Real Time Sequence Data

Real time sequence data is pathogen genome data that is time-resolved and geo-located. This technique helps us to understand and respond to the evolutionary development of pathogens. Sequence data can help track the emergence of drug resistance, and it can be combined with other sources of data to perform transmission network analysis. Moreover, sequencing can provide insights about transmission across species, as well as the spread across space and time. An effective analysis requires a large sample size on the order of hundreds or thousands. The technique has been used to assess HIV and influenza (Woolhouse et al. 2015). The 2009 H1N1 flu pandemic was the first major outbreak to be tracked in real time with virus genetic data. Since then, it has been used to track the 2014 Ebola outbreak in West Africa (Stadler et al. 2014) and COVID-19 (CDC 2020).

During the 2019 H1N1 pandemic, researchers relied on data from GISAID's EpiFlu database for real time sequence data. Launched in 2008, the EpiFlu database allowed countries to publicly share genetic data and track the evolution of the virus as it spread (Shu and McCauley 2017). Within 2 months of the first cases, researchers were able to provide accurate estimates of the evolutionary rate, date of emergence, and transmission rate using real time genetic sequencing, allowing for early characterization of the epidemic (Hedge et al. 2013).

Box 4: Using Real Time Genome Data to Counter Popular Misconceptions During COVID-19

In studying the genome data of COVID-19 patients in New York City, researchers found that most of the cases were introduced from Europe and also found evidence for community spread. This countered the common belief that the virus came to the United States from China; a misconception that had informed the policy decision to shut down the border with China much earlier than that of Europe. If sequence data had been used in more real time during the early days of COVID-19 in the United States, it could have informed these border policies and may have prevented further introduction of the disease by travelers coming from Europe (Gonzalez-Reiche et al. 2020).

Syndromic Data

Syndromic data is a type of early warning system for disease outbreaks. While the type of data used by syndromic surveillance systems can vary, it centers around a systematic approach in which health departments use automated data acquisitions and alerts to monitor disease indicators as close to real time as possible (Henning 2004). These indicators can include everything from school absenteeism and increased over-the-counter medication sales to veterinary data that reveals an unexpected increase in animal deaths (Abat et al. 2016). Syndromic surveillance focuses on the early symptom periods of a disease, before clinical or lab tests can traditionally provide a diagnosis, allowing for earlier detection of outbreaks. Negatively, these alternative, non-clinical, data sources can have confounding factors that show spikes unrelated to disease outbreaks, but when combined with multiple sources, they can provide a more reliable picture (Henning 2004).

A study in Madhya Pradesh, India stationed data collectors in clinics using specially designed mobile apps to collect demographic information and symptomatic data. The data was then submitted to a central server for analysis. By using cell phones, the data collectors were able to obtain syndromic data about rural India that hadn't been previously accessible. The cell phone collection gave researchers access to data in electronic form in order to improve syndromic surveillance compared with manually entered data (Diwan et al. 2015).

Non-Health Data Relevant for Disease Monitoring

While disease surveillance historically relies on health data collected through the formal health system, there are other sources of related data that offer insights. Digital big data, satellite data and telecommunications data in particular, are becoming more and more accessible for public health reporting, catalyzed by the data revolution for sustainable development (IEAG 2016).

Non-Health Digital Data

The growing volumes of digital data now available are a resource for real time insights about a range of social conditions, including the spread of disease. The earliest applications of digital data for disease surveillance drew on web search data to estimate the prevalence of influenza within certain communities. This method was made famous by Google Flu Trends. However, Google Flu Trends failed to detect 2009 flu outbreak (Woolhouse et al. 2015) and the tool was ended in 2015.

There is now more interest in using social media data rather than search data, but flu is still the most commonly monitored disease (Paul et al. 2016). With unprecedented numbers of users on social media (as of January 2020, Twitter had 330 million active monthly users, most outside of the USA), there are opportunities to combine this information with machine learning and natural language

processing to monitor public health (Paul et al. 2016). Search and Twitter data have been used to track dengue fever (Paul et al. 2016). Twitter has been used to track cholera, Ebola, and e. coli and has also been considered for monitoring HIV (Stoové and Pedrana 2014).

There are some difficulties when using social media data. Short text, such as Tweets, does not lend itself to natural language processing, and the colloquial language used in social media makes it challenging to search for and classify posts (Paul et al. 2016). There also needs to be a more extended consideration of the ethics involved. Although social media data may technically be public, analyzing posts can reveal private information.

Telecommunications Data

Predicting and containing the spread of a disease requires an understanding of population movement. Telecommunications data provides an innovative way for estimating such population movement, and it has been used to help address disaster scenarios, disease outbreaks, and service planning. By looking at changes in where calls were received, messages sent, and other mobile network activity, it is possible to document how groups of people may have relocated over time.

Location data from call data records were used as part of malaria prevention strategies (Wesolowski et al. 2012). Building on initial work, it was then proposed the approach be used for responding to the West Africa Ebola outbreak of 2014–2016 (Wesolowski et al. 2014). The rapid spread of Ebola was thought to be driven by local and regional travel, and mobility data could inform epidemiological models. During the Ebola outbreak, Sierra Leone created a containment strategy that was highly controversial, with many dismissing it as counterproductive. Traditional methods, such as self-reporting surveys, are not well suited for measuring the effectiveness of containment, though. Researchers accessed call data records following a lockdown event, and they found a 31% reduction in mobility for distances under 15 km and a 76% reduction for distances beyond 30 km, with original travel patterns returning after the lockdown was lifted. Furthermore, the impact was up to twice as great in areas with a higher disease burden, suggesting that less affected areas might not have been as inconvenienced. This experience shows the potential of anonymized, mobile data to capture complex population dynamics and movement, and inform future disease responses (Peak et al. 2018).

Modeling and Satellite Estimates

Mathematical models can provide real time projections of disease spread (Woolhouse et al. 2015). There are many different approaches, but disease spread is basically modeled as an exponential process, defined by the parameters of the number of secondary people a first person will infect and the generation time.

Risk mapping looks at various predictors of disease outbreak and spread, including elevation, vegetation, animal species present, other environmental factors, and so on (Woolhouse et al. 2015). Especially in remote areas, mapping requires satellite imagery and other supplemental data. Identifying areas of acute risk through mapping can help direct surveillance efforts and resources between outbreaks.

GIS and satellite images can help monitor variables like temperature, precipitation, humidity, wind, and other variables that affect the spread of diseases. As such satellite imagery has been used in studies and forecasts of diseases such as Hantavirus pulmonary syndrome (HPS), malaria, dengue, Lyme, and Rift Valley fever (Nsoesie et al. 2015). During an outbreak, analysis of spatial data can identify the spread of the disease, show what population groups are at risk, evaluate how many facilities are available to provide healthcare within a certain distance, and assess the effectiveness of control measures (Singh and Ranjan 2015).

Box 5: Using Satellite Data to Estimate Hospital Attendance
Researchers in Argentina, Chile, and Mexico used satellite imagery to monitor hospital parking lot traffic data to augment public health disease surveillance. They hypothesized that increases in hospital traffic could serve as an early indicator of social disruption resulting from disease. They used high-resolution satellite imagery collected from January 2010 to May 2013 and overlaid it with data on the incidence of respiratory virus illnesses, collected by the Pan American Health Organization (PAHO). They then developed dynamical Elastic Net multivariable linear regression models to estimate the incidence of respiratory virus illnesses using hospital traffic. The models for influenza and other respiratory viruses using hospital traffic data for select hospitals in Chile, Argentina and Mexico, performed well 'in capturing the trends present in the data within a reasonable range of error.' The errors were partly explained by lags in data releases from Ministries of Health, as well as PAHO. The researchers also concluded the model could not properly account for high-density parking (i.e., multistory car parks) or for socialeconomic dynamics relating to car ownership, shared usage, and so on. Nonetheless, the project demonstrated that, when combined with other information, this type of satellite data could be useful in monitoring disease trends (Nsoesie et al. 2015).

New Data Limitations

New data sources provide huge possibility for monitoring health conditions and disease transmission, not least of all satellite and telecommunications data which can help us to map population movement and disease spread in real time. However, as alluded to above, each approach has its own pitfalls and is best utilized in combination with other traditional, and potentially more representative, methods. Common

limitations affecting all of these methodologies are insufficient capacity, representative bias, ethical concerns and the complexity of data-sharing partnerships.

Insufficient Diagnostic Testing and Long-Term Capacity Development

As discussed above (2.2) diagnostics are a fundamental component of successful outbreak containment or control strategies. And we've seen huge advances in rapid diagnostic approaches and techniques. However, as demonstrated by COVID-19, but also recent Ebola, Zika, and yellow fever outbreaks, there are common barriers to diagnostic preparedness that occur across all epidemic situations.

> In the case of the 2013–2016 Ebola epidemic in West Africa, there was a 3-month delay between the index case and the identification of the causative agent; post-outbreak analyses suggest that diagnosing 60% of patients within 1 day instead of 5 days could have reduced the attack rate from 80% to nearly 0% (Haug et al. 2016 in Kelly-Cirino et al. 2019)

The primary limitation is inadequate diagnostic testing capacity at both national and community levels of healthcare, with governments loathe to invest in pandemic preparedness outside of crisis periods. Furthermore, the companies producing the diagnostic testing for pathogens need incentives to work on these outside of outbreak periods (Kelly-Cirino et al. 2019). The problem with this is that test development and validation, through field tests, only commence when the pathogen has substantively taken hold, slowing down the whole diagnostic process.

These challenges are not unique to diagnostics. Few governments are investing substantially in broad national or subnational data and monitoring capacities, particularly using new or alternative techniques such as GIS, telecommunications analytics and so on, meaning national and subnational health actors, and statisticians, are ill-equipped to use these approaches during the crisis. However, with the frequency of pandemics increasing this kind of testing and capacity development is essential to ensure the ongoing health of the population.

For diagnostics, other long-term investments should focus on identifying overlaps in diagnostic development needs across different priority pathogens, which would prove timelier and more cost-effective than a pathogen by pathogen-based approach (Kelly-Cirino et al. 2019).

Representativeness

Another issue affecting many of the new data approaches is representativeness. Telecommunications data only provides demographic data on mobile phone users and does not differentiate between shared mobile phone usage. Digital data, such as twitter analysis, only captures the digitally connected and those actively engaged in social media (Bansal et al. 2016). And GIS data infers population characteristics

based on other proxies such as infrastructure and roofing. It is not a representative, house to house count. As such all of these methods require triangulation with censuses, household surveys and other representative data collection sources to ensure demographic accuracy.

For example, researchers have raised questions about the representativeness of mobile data used as part of the Ebola response. Anthropological studies have shown that cell phones are not used strictly as individual property in West Africa, meaning that they are not reliable beacons of individual behavior (Erikson 2018). Instead, ownership is much more fluid, with one phone possibly being shared by multiple family members, and with one individual potentially owning multiple phones. Furthermore, with patchy phone service in less urban areas and many phones not having GPS capabilities, it can be challenging to accurately locate a call. On top of these technical limitations, Ebola is spread by person to person contact, so the existing modeling assumptions for malaria that underpinned the analysis of mobile data might not translate well to Ebola. Critics also point out that there has not yet been much in the way of peer reviews of the positive claims about telecommunications data for surveillance (Maxmen 2019). And indeed, to prevent Ebola transmission, restricting overall movement is less important than isolating patients to prevent direct contact with infected fluids (Maxmen 2019). These concerns emphasize the need to fully consider the local and social context as well as the specifics of a disease before deploying novel data solutions.

Ethical Concerns

The use of new data sources has raised a variety of ethical considerations that must be balanced with the benefits to public health, for example, the concern that much public health information is becoming concentrated with select, powerful companies (Kostkova 2018). As such, there is an urgent need for guidelines around the acceptable use of personal data (by both the public and the private sector) while also providing sufficient flexibility to accommodate different emerging data sources and respecting issues of global justice (Vayena et al. 2015). Already, there are growing concerns about the lack of consent involved with the sharing of data for public health applications, as well as about the fairness of algorithms built off of these data (McDonald 2016). Because of the high-resolution of spatial data used for disease surveillance, privacy is another important concern. Current safeguards focus on anonymizing individually identifiable data and the aggregation of shared data to protect privacy. However, even when data is anonymized, researchers have shown that it can be de-anonymized with very little information, enabling them to uniquely identify individuals represented in the dataset (Kondor et al. 2018).

Ensuring the methodological robustness of new techniques can also be viewed as an ethical issue, not just a scientific one (Vayena et al. 2015). Epidemiological studies are prone to a wide range of methodological challenges (CDC 2006b), which may incorrectly estimate the spread of a disease leaving certain individuals and communities vulnerable to health risks, while at the same time economically

harming, stigmatizing, or curtailing the freedoms of others without necessarily creating public health benefits.

This set of ethical concerns will need to be addressed with concerted action. Privacy issues could be mitigated in part with higher levels of aggregation, lower spatial resolution, and the use of synthetic data sets (Kondor et al. 2018). Although this creates a potential tradeoff between the granularity of data and privacy protection, public health experts should consider the minimum level of granularity needed to conduct epidemiological analysis (as noted by Public Health England 2012). Regarding the wider set of concerns, it has been suggested that the epidemiological community develop best-practice standards, such as review boards to assess the potential risks and benefits for communities and to negotiate compensation in the event of accidental harm (Vayena et al. 2015).

Negotiating Partnerships

Much of the innovations discussed above require collaboration between the public sector and private or third-party actors, and as such require a formal partnership or data sharing agreement. These agreements are important so that new sources of data can be validated and then integrated into traditional statistical and monitoring systems (Bansal et al. 2016). They are also important to ensure private data is maintained and can be relied upon over the medium to long term for national reporting. As mentioned above, Google Flu received widespread attention for its apparent ability to track the spread of the flu in real time through web search queries, but the tool eventually lost its efficacy and was discontinued. This example underscores the issue of volatility when data is not created for the direct purpose of public disease surveillance and there is less incentive to maintain it (Bansal et al. 2016).

However, negotiating partnership agreements can be difficult and time consuming. A consortium of actors including SDSN TReNDS, NYU Gov Lab, University of Washington, and The World Economic Forum, analyzed a host of data sharing agreements and found serious inequalities in the partnership terms, very unequal capacities to negotiate the partnerships, challenges relating to citizen's privacy and so on (Dahmm 2020). In summary, they highlight six important issues that need to be considered when embarking on data sharing partnerships; (1) Why the data is bring shared? (2) What kinds of data are being shared? (3)When the data should be shared? (4) Who is involved in the data sharing, (5) How the data is being shared, and (6) Where the data is being shared from and to?

As we enter an era of more complex and multi-party data production, ownership and sharing, formalized data partnerships will be crucial but will require institutional innovations. Governments will need to designate partnership brokers, ensure adequate legal capacity, and familiarity with partnership terms. Likewise, private data actors will need to ensure their practices comply with national standards for data ownership, control and use, if they are to maintain a long-term partnership with any public authority.

National Systems Innovation

Disease surveillance data, whether traditional or new, needs to be incorporated into a system that is able to appropriately assess and respond to risk.

In the early 2000s, Europe advanced the notion of Epidemic Intelligence (EI), which covers all activities for identifying new health risks (Paquet et al. 2006). This includes both structured data from routine surveillance (or "indicator surveillance") and unstructured data from any source of intelligence more generally (or "event surveillance"). The importance of EI is now recognized globally within International Health Regulations (2005) (IHR), specifically the guidance on surveillance which aims to protect the global community from transboundary public health risks and emergencies (WHO 2005). Today a number of countries and regions around the world have advanced epidemic intelligence facilities, training, and centers (such as the UK, USA, and the European Union), but sadly this is still not the norm, with limited progress having been made in the vast majority of countries (Woolhouse et al. 2015).

There is a need for investment and capacity building, and increased connections between the different components of national surveillance systems. Data sharing can bring a proliferation of data, but it can then lead to uncertainty about the quality of methods and loss of control on the part of the National Statistics Office or Ministry of Health. In 2015, SDSN TReNDS recommended the creation of Chief Data Officers; officials responsible for coordinating new data partnerships and sources, working closely with the National Statistical Office to validate the data and integrate it into a national system (Espey et al. 2017). Such capable systems are essential to managing the variety of data tools and partners now available.

Conclusion

To fully utilize new data tools for disease surveillance and monitoring, there needs to be investment and strengthening of digital capacity. Tools should not be learned and deployed during a disaster, but should be made available and integrated in anticipation of such (Woolhouse et al. 2015), thereby also providing time for the tools to be assessed vis-a-vis local social context and vetted alongside alternative methods. Also important is to provide sufficient time to negotiate partnership agreements between national governments and third-party actors to ensure fair data sharing practices and sustainable arrangements. Above all, health indicators should be scientifically rigorous, and we should not forget the fundamental importance of traditional, bottom-up data, even amid the hype of remote sensing and digital data tools. But used together within an improved surveillance system, the discussed data innovations can help to improve the equity of public health responses and help to ensure that no one is left behind.

References

Abat C, Chaudet H, Rolain JM, Colson P, Raoult D (2016) Traditional and syndromic surveillance of infectious diseases and pathogens. Int J Infect Dis 48:22–28, ISSN 1201-9712

Arnold BF, Scobie HM, Priest JW, Lammie PJ (2018) Integrated Serologic Surveillance of Population Immunity and Disease Transmission. Emerg Infect Dis 24(7):1188–1194. https://doi.org/10.3201/eid2407.171928

Bansal S, Chowell G, Simonsen L, Vespignani A, Viboud C (2016) Big data for infectious disease surveillance and modeling. J Infect Dis 214:S375–S379. https://doi.org/10.1093/infdis/jiw400

Bell BP, Damon IK, Jernigan DB et al (2016) Overview, control strategies, and lessons learned in the CDC response to the 2014–2016 Ebola epidemic. Morb Mortal Wkly Rep 65:4–11

Broadhurst M, Brooks T, Pollock R (2016) Diagnosis of Ebola virus disease: past, present, and future. Clin Microbiol Rev 29:773–793. https://doi.org/10.1128/CMR.00003-16

Campos-Outcalt D, England R, Porter B (1991) Reporting of communicable diseases by university physicians. Public Health Rep 106:579–583

CDC (2006a) Global public health surveillance under new international health regulations. Emerg Infect Dis J 12(7). Available via CDC https://wwwnc.cdc.gov/eid/article/12/7/05-1497_article. Accessed 6 Aug 2020

CDC (2006b) Principles of epidemiology in public health practice: an introduction to applied epidemiology and biostatistics, 3rd edn. U.S. Department of Health and Human Services, Atlanta. Available via https://www.cdc.gov/csels/dsepd/ss1978/SS1978.pdf. Accessed 28 Aug 2020

CDC (2012) Principles of epidemiology in public health practice, third edition, an introduction to applied epidemiology and biostatistics. Public Health Surveillance, Lesson Five. Available via CDC https://www.cdc.gov/csels/dsepd/ss1978/index.html. Accessed 31 July 2020

CDC (2020) COVID-19 serology surveillance strategy. Centers For Disease Control and Prevention. Available via CDC https://www.cdc.gov/coronavirus/2019-ncov/covid-data/serology-surveillance/index.html. Accessed 31 July 2020

Chunara R, Aman S, Smolinski M, Brownstein JS (2013) Flu near you: an online self-reported influenza surveillance system in the USA. Online J Public Health Inform 5(1). https://doi.org/10.5210/ojphi.v5i1.4456

Dahmm H (2020) Laying the foundation for effective partnerships: an examination of data sharing agreements. SDSN TReNDS working paper, Pre-print. Available via SDSN TReNDS https://www.sdsntrends.org/research/dsainsightsreport. Accessed 6 Aug 2020

Diwan V, Agnihotri D, Hulth A (2015) Collecting syndromic surveillance data by mobile phone in rural India: implementation and feasibility. Glob Health Action 8(1):26608. https://doi.org/10.3402/gha.v8.26608

Erikson SL (2018) Cell phones as an anticipatory technology: behind the hype of big data for Ebola detection and containment. German Research Foundation working paper. Available via Social Science in Humanitarian Action Platform https://www.socialscienceinaction.org/resources/cell-phones-anticipatory-technology-behind-hype-big-data-ebola-detection-containment/. Accessed 6 Aug 2020

Espey J (2015) Data for development: a needs assessment for SDG monitoring and statistical capacity development. Working paper. Available via SDSN https://resources.unsdsn.org/data-for-development-a-needs-assessment-for-sdg-monitoring-and-statistical-capacity-development. Accessed 6 Aug 2020

Espey J, et al (2017) Counting on the world: building modern data systems for sustainable development. Worling paper. Available via SDSN TReNDS https://countingontheworld.sdsntrends.org/2017/. Accessed 15 Aug 2020

Gonzalez-Reiche AR, et al (2020) Introductions and early spread of SARS-CoV-2 in the New York City area. Pre-Print. https://doi.org/10.1101/2020.04.08.20056929

Haug CJ, Kieny MP, Murgue B (2016) The Zika challenge. N Eng J Med 374:1801–1803. https://doi.org/10.1056/NEJMp1603734

Hay SI, George DB, Moyes CL, Brownstein JS (2013) Big data opportunities for global infectious disease surveillance. PLoS Med 10(4). https://doi.org/10.1371/journal.pmed.1001413

Hedge J, Lycett SJ, Rambaut A (2013) Real-time characterization of the molecular epidemiology of an influenza pandemic. Biol Lett. https://doi.org/10.1098/rsbl.2013.0331

Henning K (2004) What is syndromic surveillance? Morb Mortal Wkly Rep 53:7–11. Available via JSTOR from www.jstor.org/stable/23315680. Accessed 13 July 2020

IEAG (2016) A world that counts: mobilising a data revolution for sustainable development. Report by the Independent Expert Advisory Group on the data revolution, for the UN Secretary-General. Available at https://www.undatarevolution.org/report/. Accessed 6 Aug 2020

IHME (2017) Findings from the global burden of disease study. Institute for Health Metrics and Evaluation. Available via HealthData.org http://www.healthdata.org/policy-report/findings-global-burden-disease-study-2017. Accessed 6 Aug 2020

IHME (2020a) Global health data exchange. Institute of Health Metrics and Evaluation. Available via Global Health Data Exchange http://ghdx.healthdata.org/gbd-results-tool. Accessed 30 July 2020

IHME (2020b) IHME: about verbal autopsy. Website page. Available via http://www.healthdata.org/verbal-autopsy/about. Accessed 28 Aug 2020

Kelly-Cirino CD, Nkengasong J, Kettler H, Tongio I, Gay-Andrieu F, Escadafal C, Piot P, Peeling RW, Gadde R, Boehme C (2019) Importance of diagnostics in epidemic and pandemic preparedness. BMJ Glob Health 4(Suppl 2):e001179. https://doi.org/10.1136/bmjgh-2018-001179

Kondor B, Hashemian Y, de Montjoye, Ratti C (2018) Towards matching user mobility traces in large-scale datasets. In: IEEE transactions on big data. https://doi.org/10.1109/TBDATA.2018.2871693

Kost G (2020) Geospatial hotspots need point-of-care strategies to stop highly infectious outbreaks: Ebola and coronavirus. Arch Pathol Lab Med. https://doi.org/10.5858/arpa.2020-0172-RA

Kostkova P (2018) Disease surveillance data sharing for public health: the next ethical frontiers. Life Sci Soc Policy 14(1):16

Kozel TR, Burnham-Marusich AR (2017) Point-of-care testing for infectious diseases: past, present, and future. J Clin Microbiol 55(8):2313–2320. https://doi.org/10.1128/jcm.00476-17

Laurie et al (2013) Influenza serological studies to inform public health action: best practices to optimise timing, quality and reporting. Influenza Other Respir Viruses 7(2):211–224. https://doi.org/10.1111/j.1750-2659.2012.0370a.x

Maxmen A (2019) Can tracking people through phone-call data improve lives? Nature 569:614–617. https://doi.org/10.1038/d41586-019-01679-5

McDonald SM (2016) Ebola: a big data disaster-privacy, property, and the law of disaster experimentation. The Centre for Internet and Society. Available via the Center for Internet Society https://cis-india.org/papers/ebola-a-big-data-disaster. Accessed 6 Aug 2020

Nsoesie EO, Butler P, Ramakrishnan N, Mekaru SR, Brownstein JS (2015) Monitoring disease trends using hospital traffic data from high resolution satellite imagery: a feasibility study. Sci Rep 5(1):x. https://doi.org/10.1038/srep09112

Paquet C, Coulombier D, Kaiser R, Ciotti M (2006) Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. Euro Surveill 11(12):212–214

Paul MJ, Sarker A, Brownstein JS, Nikfarjam A, Scotch M, Smith KL, Gonzalez G (2016) Social media mining for public health monitoring and surveillance. In: Pacific symposium on biocomputing. Available via Stanford https://psb.stanford.edu/psb-online/proceedings/psb16/intro-smm.pdf. Accessed 28 Aug 2020

Peak CM, Wesolowski A, zu Erbach-Schoenberg E, Tatem AJ, Wetter E, Lu X, Power D, Weidman-Grunewald E, Ramos S, Moritz S, Buckee CO (2018) Population mobility reductions associated with travel restrictions during the Ebola epidemic in Sierra Leone: use of mobile phone data. Int J Epidemiol 47(5):1562–1570. https://doi.org/10.1093/ije/dyy095

Pilot E, Jena B, Krafft T, Ziemann A, Rao R, Kauhl B, Pardeshi RR, Salvi S (2011) New approaches to disease surveillance and early outbreak detection in India. Proc Fourth Int Conf Health GIS 5:6

Public Health England (2012) Public health surveillance: towards a public health surveillance strategy for England. Department of Health online working paper. Available via UK Government https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/213339/Towards-a-Public-Health-Surveillance-Strategy.pdf. Accessed 28 Aug 2020

Shu Y, McCauley J (2017) GISAID: global initiative on sharing all influenza data – from vision to reality. Euro Surveill. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Singh R, Ranjan K (2015) Satellite imaging and surveillance of infectious diseases. J Trop Dis. https://doi.org/10.4172/2329-891x.1000s1-004

Soucie JM (2012) Public health surveillance and data collection: general principles and impact on hemophilia care. Hematology (Amsterdam, Netherlands) 17 Suppl 1(0 1):S144–S146. https://doi.org/10.1179/102453312X13336169156537

Stadler T, Kühnert D, Rasmussen DA, du Plessis L (2014) Insights into the early epidemic spread of ebola in Sierra Leone provided by viral sequence data. PLoS Curr 6. https://doi.org/10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f

Stoové MA, Pedrana AE (2014) Making the most of a brave new world: opportunities and considerations for using Twitter as a public health monitoring tool. Prev Med 63:109–111. https://doi.org/10.1016/j.ypmed.2014.03.008

Streatfield PK, Khan WA, Bhuiya A (2014) HIV/AIDS-related mortality in Africa and Asia: evidence from INDEPTH Health and Demographic Surveillance System sites. Glob Health Action 7: 25370

United Nations Environment Programme (2020) Preventing the next pandemic – Zoonotic diseases and how to break the chain of transmission. United Nations Environment Programme. Available via UNEP https://www.unenvironment.org/resources/report/preventing-future-zoonotic-disease-outbreaks-protecting-environment-animals-and. Accessed 30 July 2020

Vayena E, Salathé M, Madoff LC, Brownstein JS (2015) Ethical challenges of big data in public health. PLoS Comput Biol 11(2):e1003904

Watson C, Cicero A, Blumenstock J, Fraser M (2020) A National Plan to Enable Comprehensive COVID-19 Case Finding and Contact Tracing in the US, Johns Hopkins Bloomberg School of Public Health. Johns Hopkins University: Baltimore. https://www.centerforhealthsecurity.org/our-work/pubs_archive/pubs-pdfs/2020/200410-national-plan-to-contact-tracing.pdf

Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science. https://doi.org/10.1126/science.1223467

Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ (2014) Commentary: containing the Ebola outbreak-the potential and challenge of mobile network data. PLoS Curr Outbreaks 1. https://doi.org/10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e

WHO Ebola Response Team, Aylward B, Barboza P, Bawo L, Bertherat E, Bilivogui P, Blake I, Brennan R, Briand S, Chakauya JM, Chitala K, Conteh RM, Cori A, Croisier A, Dangou JM, Diallo B, Donnelly CA, Dye C, Eckmanns T, Ferguson NM et al (2014) Ebola virus disease in West Africa – the first 9 months of the epidemic and forward projections. N Engl J Med 371 (16):1481–1495. https://doi.org/10.1056/NEJMoa1411100. Accessed 28 August 2020

Woolhouse ME, Rambaut A, Kellam P (2015) Lessons from Ebola: improving infectious disease surveillance to inform outbreak management. Sci Transl Med 7(307):307rv5. https://doi.org/10.1126/scitranslmed.aab0191

World Bank and World Health Organization (2014) Global civil registration and vital statistics: scaling up investment plan 2015–2024. Available via WHO https://www.who.int/healthinfo/civil_registration/WB-WHO_ScalingUp_InvestmentPlan_2015_2024.pdf?ua=1. Accessed 31 July 2020

World Health Organisation (2005) Strengthening health security by implementing the International Health Regulations. WHO Online. Available at https://www.who.int/ihr/surveillance/en/. Accessed 6 Aug 2020

World Health Organisation (2017) Verbal autopsy standards: the 2016 WHO verbal autopsy instrument, V1.5. Available at https://www.who.int/healthinfo/statistics/verbalautopsystandards/en/. Accessed 6 Aug 2020

World Health Organization (1999) WHO recommended disease surveillance standards – second addition. World Health Organization. Accessible via WHO https://www.who.int/csr/resources/publications/surveillance/whocdscsrisr992.pdf. Accessed 6 Aug 2020. Accessed 28 Aug 2020

World Health Organization (2020) Antimicrobial resistance. World Health Organization. Available via WHO https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance. Accessed 6 July 2020

World Pop (2020) Percentage of women of childbearing age living within two hours of a hospital. Available via SDGs Today https://www.arcgis.com/home/item.html?id=30756d51b0ae4a2dad49f16b33920750. Accessed 6 Aug 2020

Ye Y, Wamukoya M, Ezeh A, Emina JB, Sankoh O (2012) Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa. BMC Public Health 12:741. https://doi.org/10.1186/1471-2458-12-741