Tracking Disease

Digital epidemiology offers new promise in predicting outbreaks.

By Mary Bates

On 19 October 2010, ten months after a devastating earthquake hit Haiti, the Haitian Ministry of Public Health and Population (MSPP) was notified of a sudden surge in patients suffering from watery diarrhea and dehydration. Two days later, the Haiti National Public Health Laboratory identified the culprit: *Vibrio cholerae*. On 22 October, officials announced the first cholera outbreak in Haiti in more than a century.

The MSPP, with technical assistance from the U.S. Centers for Disease Control and Prevention (CDC) and its partners, employed many traditional forms of epidemiological surveillance to track the outbreak and try to stem its spread. These included visiting hospitals, interviewing patients in the communities where the disease was first reported, conducting telephone interviews with health facilities, and reporting the results to the press.

But they also turned to some nontraditional surveillance methods. Because 86% of the world's population lives under mobile cellular network coverage and mobile phone networks routinely register data that can be used to track the location of all active mobile phone users, researchers were able to track the movement of people in Haiti via their mobile phones, a fact that would become critical in predicting the spread of the disease. Using position data from subscriber identity module cards provided by the largest mobile phone company in Haiti, researchers estimated the magnitude and trends of population movements, allowing for close to real-time monitoring of the disease outbreak. In fact, studies show that Twitter data following the 2010 earthquake accurately tracked the cholera spread, and these data were faster and roughly as correct as the official records detecting the beginning and early progress of the epidemic.

Advances in automated data processing and machine learning now allow epidemiologists to meticulously sift through the millions of digital traces we collectively leave behind each day

Digital Object Identifier 10.1109/MPUL.2016.2627238
Date of publication: 20 January 2017

©ISTOCKPHOTO.COM/BAKHTIAR_ZEIN

as we conduct our lives online—through Internet searches, social media posts, or the use of our mobile phones. There is a rich treasure trove of data that, though it may provide only a partial glimpse of a disease outbreak, is proving to be quite nimble at capturing the spread of disease. Traditional tracking methods—surveillance of clinics, telephone triage calls, purchases of over-the-counter medications, and school absenteeism, for example—are far more unwieldy. Although researchers caution that these new sources of information cannot completely replace traditional epidemiological methods, they are serving to fill in critical gaps.

"The epidemiological landscape is rapidly changing, primarily because of this rapid diffusion of new technologies," says Georgia Tourassi, director of the Biomedical Science and Engineering Center at Oak Ridge National Laboratory in Tennessee (Figure 1). "The Internet, mobile technology, social media, all these technologies," she continues, "allow us to collect information at a potentially much higher resolution than we could before."

Tracking Outbreaks Through Searches and Social Media

Web-search queries were one of the first digital data sources to be used in disease surveillance. According to Harvard Medical School epidemiologist John Brownstein, an estimated 37–52% of Americans seek health-related information on the Internet each year, mainly using search engines to find details about symptoms and treatments. Scientists can analyze users' search terms, along with the location information encoded in their computers' Internet protocol addresses, for insights into current disease trends.

One of the earliest examples was Google Flu Trends, which began offering real-time data to the public in 2008. Based on people's Internet searches for flu-related terms, this tool monitored flu outbreaks worldwide. Although it was initially hailed as providing data nearly as precise as the CDC's while having the advantage of being one to two weeks faster, Google shut down Flu Trends in 2015 after some of its estimates proved inaccurate. Despite this early setback, Google records are still useful epidemiologic tools. Google now sends its search data to scientists at the CDC, Columbia University, and Boston Children's Hospital, where scientists are developing more refined algorithms to mine the search data.

In one study, researchers from Google and the CDC used an automated method of analyzing large numbers of Google search queries to track the flu. They monitored millions of

FIGURE 1 Georgia Tourassi, director of the Biomedical Science and Engineering Center at Oak Ridge National Laboratory.

Google searches for flu-related terms and then correlated those searches to the percentage of physician visits during which a patient displayed flu-like symptoms. This method was able to consistently estimate influenza activity in each region of the United States one to two weeks ahead of the publication of CDC reports.

Social media platforms are also useful epidemiological data sources. As seen in Haiti, Twitter is quickly becoming a favorite social media data source for epidemiologists because it can provide more contextual information than search queries. "Frequently, people who are experiencing symptoms or have been confirmed to have a specific disease post about it on social media," says Mauricio Santillana, a scientist at Harvard University and Boston Children's Hospital, who has used data sources like Google and Twitter to predict disease outbreaks (Figure 2). "If we monitor the activity of tweets that contain health information," he explains, "sometimes we can spot specific outbreaks or monitor how outbreaks are developing."

In a study of the 2009 H1N1 (or swine flu) outbreak, scientists found that estimates of influenza-like illness derived from Twitter accurately tracked reported disease levels. They showed that not only could Twitter be used to estimate disease activity in real time; it could also track rapidly evolving public sentiment with respect to H1N1.

Mining Web Data for Clues

With so much digital information available, a new generation of disease-surveillance mash-ups (or web application hybrids) has arisen that can mine, categorize, filter, and visualize online intelligence about health threats and the spread of disease, often in real time.

One of these is HealthMap (Figure 3), an openly available public health intelligence system that tries to pinpoint global outbreaks in real time. Established in 2006 by a team of researchers, epidemiologists, and software developers at Boston Children's Hospital, HealthMap searches disparate online sources for reports from local news articles, witness accounts, blogs, Twitter, and official reports from the CDC and World Health Organization (WHO) to produce a global picture of ongoing infectious disease threats. Through a constantly updating, automated process, the system monitors, organizes, integrates, filters, and disseminates online information about emerging diseases. More than 100,000 people have downloaded the related mobile app, Outbreaks Near Me, which users rely on via global positioning to help them avoid infectious disease outbreaks.

HealthMap is used by a diverse audience, ranging from international travelers to local health departments to the CDC and WHO. It is useful as an early detection system as well as for providing current, highly local information about outbreaks around the world. In March 2014, HealthMap tracked early press and social media reports of Ebola in West Africa, before the WHO identified the epidemic. The HealthMap team subsequently created a visualization of the Ebola epidemic at healthmap.org/ebola. "HealthMap is a website that is constantly monitoring the flow of news alerts all over the world, in multiple languages and in real time," says Santillana. "It does not provide a complete picture of what's going on, but it gives you a sense of what people are reporting on the ground."

FIGURE 2 Mauricio Santillana, a researcher at Harvard Medical School and Boston Children's Hospital.

Recently, the Zika virus has been on a lot of epidemiologists'—and everyday folks'—minds. In March 2016, Google announced that its engineers would work with UNICEF and data scientists to create an open-source information platform to figure out the virus's path and better target response efforts. It will process data from different sources, such as weather and travel patterns, to build risk maps and visualize potential outbreaks.

Google announced the action after a 3,000% increase in global search interest in the virus. Ultimately, the goal

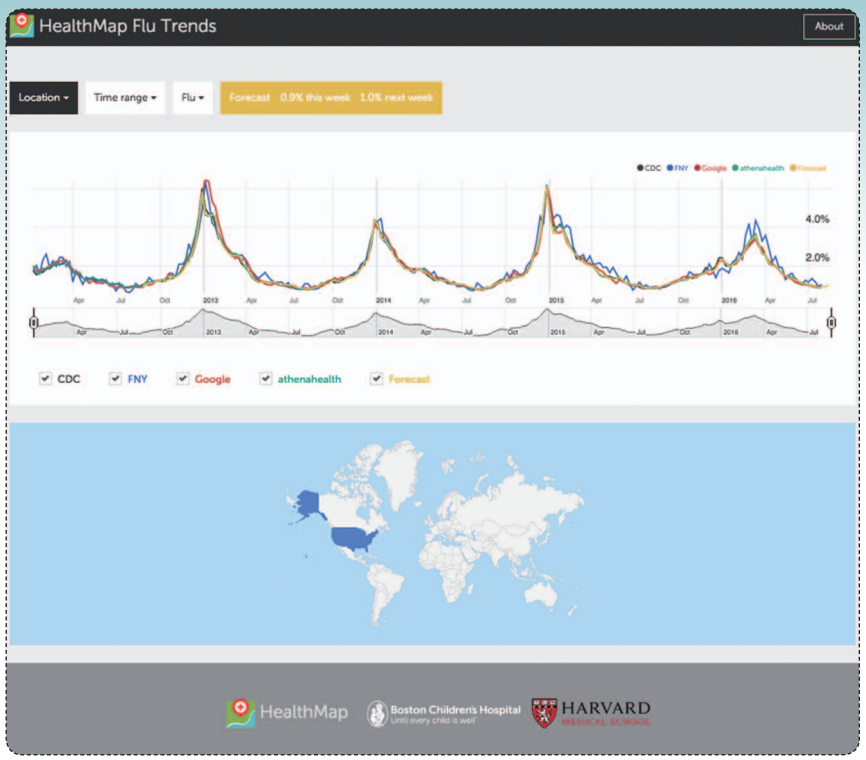

FIGURE 3 HealthMap is an open-access public health intelligence system that tracks global outbreaks, like the flu, in real time.

of the platform is to identify the risk of Zika spreading to different regions and help focus the time and resources of governments and nongovernmental organizations fighting the disease. The Zika platform is just a prototype; if successful, Google hopes it can be applied to other outbreaks.

Another open-access, web-based application, FLIRT (originally, "Flight Risk Tracker"), was successful in predicting the spread of Zika in the United States. FLIRT is a biosurveillance application, built to predict where infected travelers will likely go. It uses a database of flight schedule information from over 800 airlines that is updated monthly. With these data, FLIRT can visualize passenger flow data over flight networks.

Researchers from Michigan State University and EcoHealth Alliance used FLIRT to analyze flights departing from five selected airports in locations where a Zika outbreak had been identified. FLIRT correctly predicted which American states and cities were at the highest risk of receiving travelers infected with Zika. The results show that combining air traffic data and simulations of passenger movement creates a powerful tool to predict where infectious diseases may spread. With this information, public health officials can focus their efforts on the highest risk locations.

Combining air traffic data and simulations of passenger movement creates a powerful tool to predict where infectious diseases may spread.

The Future of Epidemiology

Web-based platforms have made disease reporting and surveillance more timely and accessible. Traditional epidemiological methods suffer from a time lag of one to two weeks, but the Internet promises more opportunities for real-time, accurate information dissemination and analysis. "The advantages are the ability to accelerate knowledge discovery and reach broader audiences at a much faster pace," says Tourassi. "These technologies are introducing a new way of reaching sectors of the population that are typically not captured by traditional systems and collecting information that otherwise may not have been accessible."

Timely, accessible, and accurate health information is especially critical for rapidly identifying outbreaks and improving public health outcomes. New digital technologies can serve as an important addition to traditional disease surveillance, especially in cases of new or emerging diseases, where little or no historical data exist. But these technologies come with several challenges.

One concern is privacy issues. Although social media data are publicly available, users may not intend or want their posts to be used for research. "When people enter search terms or post on social media, they might not know that their data [are] publicly available and may be tracked and analyzed," says Y. Tony Yang, associate professor in the Department of Health Administration and Policy at George Mason University (Figure 4). "Big companies like Google understand the privacy issue," he continues, "and in their analyses, the data cannot be traced back to the person who entered the data."

Another issue is the potentially skewed nature of such data. Internet use is heavier in big cities and among younger people, so there are concerns about just how representative of the general public the information obtained can be. However, even traditional disease-surveillance systems are prone to bias. For instance, they represent only the people who have visited a clinic or received medical attention. Web-based systems are, in many ways, reaching sections of the population not represented in traditional disease-surveillance systems.

A further concern is the accuracy of digital data. "On one hand, we have the ability to collect a lot of data, but on the other

hand, we don't have the luxury of controlling its quality," says Tourassi; "I call it big but dirty data." Santillana points to instances when there is panic in the population about a particular disease. Media reports tend to magnify the situation, and people are more likely to search for information on the Internet. "As a consequence, many web-based methods may show spikes indicating something is going on, but it may just be reflecting the panic and not the actual incidence of disease," according to Santillana.

Another challenge is that people on social media may not use accurate medical terminology. "On Twitter, for example, people don't always use standard English," says Yang. "They use more informal language, such as slang and abbreviations." A solution to this issue is the creation of special search algorithms specifically for Twitter. Researchers studying social media posts may look for text mentioning common symptoms, for instance, rather than references to specific illnesses.

All of these challenges do not mean that data from search queries or social media cannot be useful but, rather, that they shouldn't be taken on their own. The evidence so far indicates that digital data can be very meaningful, especially when compared side by side with data from traditional surveillance systems. "I see these new technologies as complementary, as augmenting traditional methods," says Santillana. "New digital technologies are promising ways to understand public health and give a more complete picture of what may be happening on the ground."

FIGURE 4 Y. Tony Yang, a professor in the Department of Health Administration and Policy at George Mason University.

Yang believes that we will be increasingly reliant on web-based surveillance in the future. "Web-based technologies are starting to make an impact," he notes. "They allow us to better respond to infectious disease outbreaks and track outbreaks in real time." Web-based technologies are still quite young, and statistical techniques for analyzing them are constantly evolving. As the Internet becomes a bigger part of everyday life for more people, it is likely that we will see web-based approaches become a more conventional form of collecting epidemiological information. "The majority of studies out there right now are more proof of concept studies to support the value of these sources," says Tourassi. "The field needs to move forward both in terms of how to capture information that is available through these new sources and how to analyze it properly, so we can aid the epidemiological community in knowledge discovery and dissemination."

As the Internet, social media, and mobile phones become bigger parts of everyday life for more people, these data sources will contribute more to epidemiological studies. We are in the midst of a digital revolution, and epidemiologists are figuring out how to mine these data for useful public health information.

Mary Bates (maryebates@gmail.com) *is a freelance science writer based in Boston, Massachusetts. Her work has been published by* National Geographic News, New Scientist, *BrainFacts.org, and other print and online publications.*