MICHELE MARCONI

BIG DATA

Cloudy with a chance of flu

Internet search data, medical records and networks of on-the-ground experts could enable the accurate forecasting and faster control of disease outbreaks.

BY MICHAEL EISENSTEIN

Even though you know it's a sensible idea, you're on the fence about whether it would be worth the bother to have this season's influenza vaccine. But a quick glance at the flu forecast on your phone sets you straight: there's a warning about a recent spike of cases nearby, so you head to the clinic rather than risk a feverish week in bed. Epidemiologists eagerly anticipate such a future, in which they can track infectious diseases with the same confidence as meteorologists mapping the weather. But those making predictions of this type face a serious problem. "There is just not a lot of observational data in the disease world," says Cécile Viboud, an epidemiologist at the US National Institutes of Health Fogarty International Center in Bethesda, Maryland. "It's several orders of magnitude less than what we have in other fields."

This year marks the centenary of the start of the Spanish flu pandemic, which involved a strain of flu virus known as H1N1 that killed up to 5% of the world's population. The world is now much better prepared for such threats, as shown by the international reaction to the H1N1 pandemic of 2009, which was coordinated by a global network of laboratories that perform clinical testing. Yet the response was not swift enough to fully contain the pandemic, which claimed the lives of about 250,000 people in the first 12 months (F. S. Dawood *et al. Lancet Infect. Dis.* **12,** 687–695; 2012).

Lawrence Madoff, an infectious-disease specialist at the University of Massachusetts Medical School in Worcester, sees such delays as being an inherent constraint of conventional lab-based surveillance strategies. "They're limited by their tendency to have rigid structures and count specific cases, and by a bureaucratic slowness that gets built into the system," he says. For example, flu surveillance in the United States relies on a network called the Influenza-like Illness Surveillance Program, through which health-care providers across the country file weekly reports of probable cases on the basis of symptoms, and submit samples from patients to testing centres. The results are assessed centrally by the US Centers for Disease Control and Prevention (CDC). Consequently, even for a well-studied disease such as flu, it can take weeks to identify and respond to an outbreak. For diseases that are not monitored routinely, the delay can be catastrophic. For example, the response to the 2014–15 outbreak of Ebola in West Africa was described by an international panel of public-health specialists as an "egregious failure", owing to the months-long delay before the World Health Organization (WHO) moved to contain what was already a full-blown emergency.

The good news is that the present era of widespread access to the Internet and digital health has created a rich reservoir of valuable data for researchers to dive into. "You could start to harness all this data that's being generated on the web, gathered across different sources, to understand population health patterns," says John Brownstein, a computational epidemiologist and chief innovation officer at Boston Children's Hospital in Massachusetts. By harvesting and combining these streams of big data with conventional ways of monitoring infectious diseases, the public-health community could gain fresh powers to catch and curb emerging outbreaks before they rage out of control.

GOING VIRAL

Data scientists at Google were the first to make a major splash using data gathered online to track infectious diseases. The Google Flu

Trends algorithm, launched in November 2008, combed through hundreds of billions of users' queries on the popular search engine to look for small increases in flu-related terms such as symptoms or vaccine availability. Initial data suggested that Google Flu Trends could accurately map the incidence of flu with a lag of roughly one day. "It was a very exciting use of these data for the purpose of public health," says Brownstein. "It really did start a whole revolution and new field of work in query data."

Unfortunately, Google Flu Trends faltered when it mattered the most, completely missing the onset in April 2009 of the H1N1 pandemic. The algorithm also ran into trouble later on in the pandemic. It had been trained against seasonal fluctuations of flu, says Viboud, but people's behaviour changed in the wake of panic fuelled by media reports — and that threw off Google's data. "Before, only people who had flu were searching for flu symptoms," says Nicholas Generous, a biosurveillance researcher at Los Alamos National Laboratory in New Mexico. "All of a sudden, people that didn't have flu were searching and that ended up giving a false result." The project never recovered because "people at Google felt that it was not worth trying to improve the algorithm", says Viboud. The company stopped supporting Google Flu Trends in August 2015, although it continued to furnish academic and governmental organizations with relevant search data. "Google was a trailblazer, but monitoring diseases was not its primary purpose," says Viboud.

Nevertheless, its work with Internet usage data was inspirational for infectious-disease researchers. A subsequent study from a team led by Cecilia Marques-Toledo at the Federal University of Minas Gerais in Belo Horizonte, Brazil, used Twitter to get high-resolution data on the spread of dengue fever in the country. The researchers could quickly map new cases to specific cities and even predict where the disease might spread to next (C. A. Marques-Toledo *et al. PLoS Negl. Trop. Dis.* **11,** e0005729; 2017). Similarly, Brownstein and his colleagues were able to use search data from Google and Twitter to project the spread of Zika virus in Latin America several weeks before formal outbreak declarations were made by public-health officials. Both Internet services are used widely, which makes them data-rich resources. But they are also proprietary systems for which access to data is controlled by a third party; for that reason, Generous and his colleagues have opted instead to make use of search data from Wikipedia, which is open source. "You can get the access logs, and how many people are viewing articles, which serves as a pretty good proxy for search interest," he says.

However, the problems that sank Google Flu Trends still exist. "Internet data is really great for diseases that are seasonal, where a lot of people get sick and there isn't a lot of media hype," says Generous. "It probably wouldn't work for Ebola." He also notes that there are challenges in interpreting how people engage with the Internet on infectious diseases: they might be worried about their own symptoms, but could also have concerns about friends or family in high-risk areas, or simply be curious. Additionally, online activity differs for infectious conditions with a social stigma such as syphilis or AIDS, because people who are or might be affected are more likely to be concerned about privacy. Appropriate search-term selection is essential: Generous notes that initial attempts to track flu on Twitter were confounded by irrelevant tweets about 'Bieber fever' — a decidedly non-fatal condition affecting fans of Canadian pop star Justin Bieber.

Alternatively, researchers can go straight to the source — by using smartphone apps to ask people directly about their health. Brownstein's team has partnered with the Skoll Global Threats Fund to develop an app called Flu Near You, through which users can voluntarily report symptoms of infection and other information. "You get more detailed demographics about age and gender and vaccination status — things that you can't get from other sources," says Brownstein. Ten European Union member states are involved in a similar surveillance programme known as Influenzanet, which has generally maintained 30,000–40,000 active users for seven consecutive flu seasons. These voluntary reporting systems are particularly useful for diseases such as flu, for which many people do not bother going to the doctor — although it can be hard to persuade people to participate for no immediate benefit, says Brownstein. "But we still get a good signal from the people that are willing to be a part of this."

"Internet data is really great for diseases that are seasonal, where a lot of people get sick."

NETWORK NEWS

Internet activity and even self-reported data still leave a lot of room for interpretation. But front-line media reports can offer more trustworthy data points for signals of infectious diseases. One of the earliest forays into online epidemiology was ProMED-mail, established in 1994 as a mailing list for public-health experts to share reports of infectious diseases — including news stories, public-health announcements and clinical observations — from around the world. ProMED-mail blossomed rapidly into a widely used service that is managed by the International Society for Infectious Diseases in Brookline, Massachusetts. "More than 70,000 people now use it, and it's become a much more organized system of moderated reports," says Madoff, ProMED-mail's editor.

The service has also spawned a more extensive effort known as HealthMap, an online atlas of infectious-disease reports built by Brownstein and his colleagues that pulls in data from ProMED-mail, reports from organizations such as the WHO and online news aggregated by Google and its Chinese counterpart Baidu. "All these news sites are out there," says Brownstein. "If you can just organize them, you can do an even better job of bringing down the time required to understand when a disease is unfolding." HealthMap extracts data automatically from these sources in real-time, giving it the advantage of speed in terms of catching a signal. But as with other attempts to computationally filter data from the Internet, researchers must be cautious of false positives such as mistaking news of malaria-related research for actual outbreaks of the disease. Accordingly, Madoff favours manual oversight for ProMED-mail. "Everything is hand-curated," he says.

When used properly, these Internet data streams can give the public-health community a head start in mobilizing a response to an outbreak. Madoff notes that ProMED-mail has pushed a number of emerging diseases into the public eye, compelling governments to take action. "We were first to report on MERS [Middle East respiratory syndrome] in Saudi Arabia in 2012," he says. "The Saudi Ministry of Health quickly responded and told us they knew about it and had a couple of other cases, and gave us more formal verification." A similar scenario played out for the outbreak of severe acute respiratory syndrome (SARS) in 2003, in which the Chinese government was initially reluctant to acknowledge the threat. "Once it was made public, they were ready to respond and became much more transparent," says Madoff. As a result, the international research community could begin to develop vaccines and treatments. And for known threats, HealthMap has outpaced conventional surveillance platforms in identifying recent infectious-disease events, including the outbreaks of both the H1N1 flu strain and Ebola. "We've shown these sources can bring down the time of detection by days or even weeks," says Brownstein.

TRUST, BUT VERIFY

All strategies for the indirect surveillance of disease still need to be clinically validated. This puts digital epidemiologists back under the constraints of conventional lab-based surveillance, which means weeks of delay — while patients seek medical care and samples are tested — before researchers can validate their signal. For less common diseases, this can be especially problematic. "The traditional surveillance data often is not there, or it's there but very patchy," says Viboud. Researchers are therefore looking to data from medicine's front line that are more reliable indicators of infectious-disease events.

Working directly with medical records is

Nicholas Generous and colleagues review data as part of efforts to forecast the spread of dengue fever.

LOS ALAMOS NATIONAL LABORATORY

one potential solution. In 2014, Viboud and her colleagues collaborated with medical-data company IMS Health (now part of Durham, North Carolina-based IQVIA), which provided de-identified medical claims filed across the United States. Analysis of the documents, which clinicians must submit to obtain reimbursement from health-insurance companies, produced a weekly view of flu transmission in US cities that was more detailed than the state-level data that are normally reported. “Medical claims are very solid because they’re based on actual visits to practitioners,” says Viboud. “To me, this is the most high-resolution data set out there.” However, such reports are also affected by delays, with many doctors filing claims weeks after seeing a patient, which makes the process better suited to post hoc epidemic analysis than to real-time surveillance.

Electronic health records could prove more useful as timely indicators of an outbreak, by helping to catch cases at the time of diagnosis. But using them poses privacy challenges — and in the United States, these data are under the control of private entities rather than a government agency, making it trickier to negotiate access. “They probably won’t make that data available to researchers — it will probably just be available to public-health officials,” says Generous. That would require the relevant local or national public-health agencies to act as intermediaries in processing and distributing health-record-derived insights to researchers, who can then use them in the modelling and analysis of epidemics.

This approach is limited to nations whose health-care systems are highly digitized, which is generally not the case for the low-income countries that have the highest burden of infectious disease. Madoff and his colleagues are trying to address this challenge through a programme called EpiCore — an army of epidemiologists with Internet access that can be mobilized to confirm reports of infection directly. “We have over 2,000 volunteer epidemiologists now in around 140 countries who agree to be contacted in the event that there is an outbreak somewhere and to try to verify it,” says Madoff. “They can do so through an online platform that allows them to remain unidentified, so we can help people who are fearful of a government crackdown or something like that.”

For now, such diagnoses are being made the old-fashioned way, with health-care workers dispatching blood and other samples from people with symptoms to dedicated labs for testing. However, rapid strides in DNA-sequencing technology are making it feasible to achieve the accurate, on-site identification of pathogenic agents at minimal cost. Soon, it could be common for mobile diagnostic labs to acquire and upload genome data in the field. In 2016, for example, an international team of researchers took to the back roads of Brazil with a low-cost, portable sequencing system developed by Oxford Nanopore Technologies, based in the United Kingdom, which enabled them to analyse samples from across northeastern Brazil at the height of the Zika virus outbreak. Such an approach could tell public-health researchers exactly which strain of pathogen they’re grappling with, as well as help them to reconstruct chains of transmission — valuable information for containing and controlling infectious diseases. “We’ll get direct viral confirmation,” says Brownstein. “I’m not sure how long that’s going to take, but it will definitely replace what we’ve been doing up until now.”

PANDEMIC-PROOF

Another challenge will be to move beyond one-off demonstrations based on single data streams to a proven system that integrates several sources of data — for example, coupling early warnings from ‘noisy’ social-media data with high-confidence signs of infection gleaned from hospital records — and that can be trained to pick up signals for several diseases at a time. “You might have a little bit of laboratory or clinical data that you can mix with Google Trends data or participatory surveillance,” says Viboud. “That’s where the field is going.” Veterinary data will also become an important piece of the puzzle, with the potential to give researchers warnings of emerging pathogens, and Madoff notes that ProMED-mail has included disease reports from livestock and wildlife since its inception. “You have to keep an eye on other species to know what might happen next in humans,” he says.

Tying these parts together will be difficult, not just because the various data sets quantify information at different scales of space and time, but also because nobody knows which combinations will improve public health. “If people actually do implement this operationally, what are the cost savings and life savings and health savings?” asks Generous. “Right now, it’s still a bit of throwing in everything but the kitchen sink and seeing what happens.” Without clear signs of their value to public health, such forays into digital epidemiology are likely to remain little more than intriguing experiments.

Yet the early evidence strongly suggests that for at least a handful of well-studied diseases, clever wrangling of data can buy the medical community extra days or weeks in which to act — time that could be used to quarantine unwell people or to mobilize clinicians or vaccine stocks. Public-health authorities are taking the idea seriously — since 2013, the CDC has run the ‘Predict the Influenza Season Challenge’ to stimulate research into outbreak forecasting. Viboud notes that the WHO has increased its focus on digital surveillance strategies following its heavily criticized response to Ebola. “The WHO hopes to get a network of modellers around the world that can help it for the next crisis,” she says.

Generous hopes that these efforts will ultimately transform into a resource for the public, enabling people to become informed consumers of epidemiological data and to take charge of their health in the same way that they might respond to information about traffic or the weather. The result could be a more sophisticated understanding of disease risk, guided by reality rather than media hype — although some education of users will be required. “When people first started to forecast weather, the idea that there was a ‘20% chance of rain’ must have been such a paradigm shift to understand, but we all get it now,” he says. “The question is, how does that happen for disease forecasting, and how does that become a routine, everyday thing?” ■

Michael Eisenstein *is a freelance science writer based in Philadelphia, Pennsylvania.*