# BIOSURVEILLANCE:
## From Text-mining to Freakidemiology

By Katharine Miller

American officials are seeking better ways to anticipate public health crises following ten years that have seen outbreaks of SARS, avian flu, H1N1, West Nile virus, cholera and, most recently, dengue fever. There's a desire to go beyond traditional disease surveillance at local, state and regional levels and find ways to deal computationally with a fire hose of potential health data. This has led to the emergence of biosurveillance systems at the intersection of epidemiology and computational methods.

November 2010 saw the publication of a new book on the topic of biosurveillance. "It's still new to the CDC and the public health world," says **Taha Kass-Hout, MS**, one of the editors of *Biosurveillance: Methods and Case Studies* and deputy director of information science in the Division of Notifiable Diseases and Healthcare Information at the Centers for Disease Control and Prevention (CDC). "A lot of things are happening now and happening fast. It's like you're flying the plane while you're building it."

Indeed, there's so much going on, it's hard to know where to start. Researchers are expanding the types of data that can be used to predict infectious disease spread; developing novel ways to analyze that data; and trying to create systems that can help address public health problems today. The approaches range from straightforward to fairly outlandish. **Philip Polgreen, MD**, assistant professor of medicine at the University of Iowa, says he tries to be different. His infectious disease work uses futures markets, data from social media sites, and iPhone apps. "I call it freakidemiology," Polgreen says.

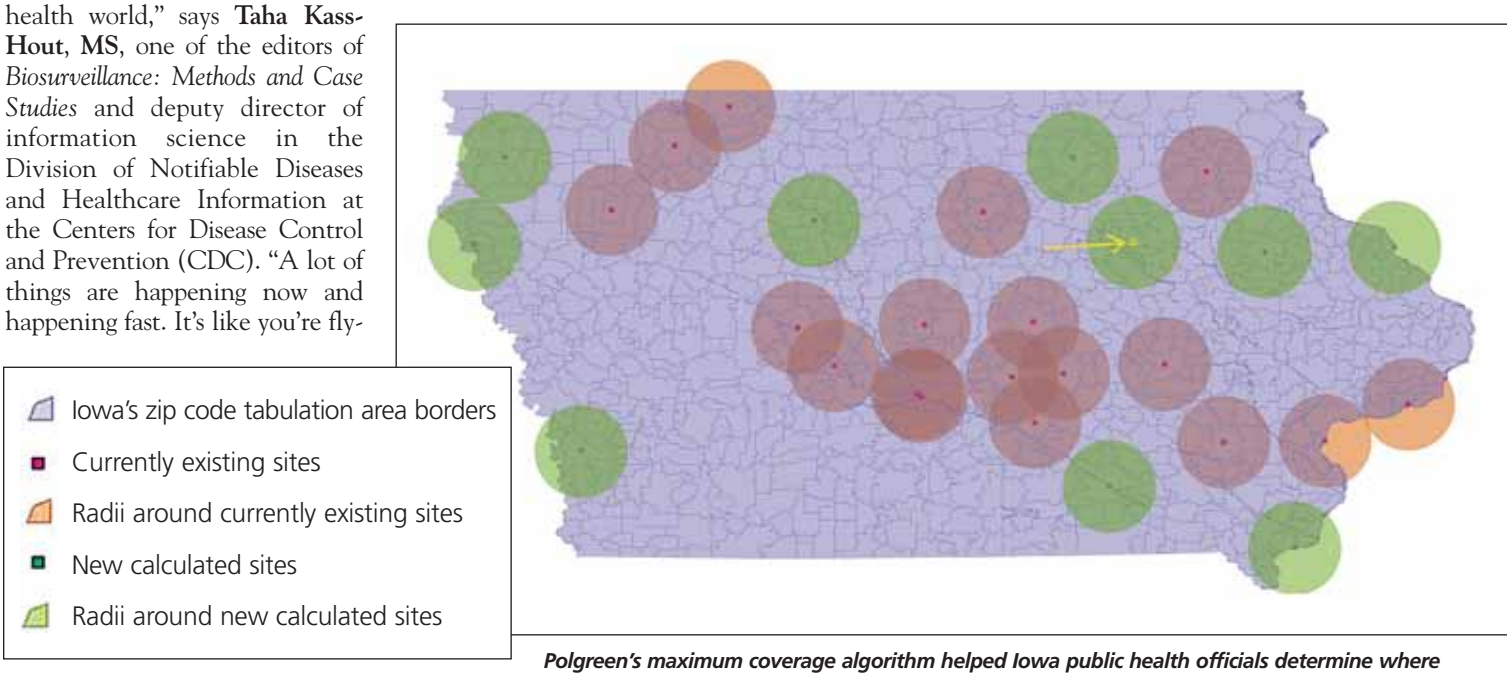
Freakish or not, biosurveillance currently provides an exciting and active niche for computational biology, with the potential to impact human health.

### Expanding on Traditional Inputs

The CDC's flagship program, called BioSense, is currently computationally straightforward: It's fed with expert rules that the system uses to understand disease complaints from national, regional, and local health data sources (e.g., clinical laboratories, health departments' syndromic surveillance systems [emergency rooms, ambulatory care sites], US Department of Defense and Veterans Administration medical treatment facilities, and pharmacy chains). But according to Kass-Hout, "the next generation of biosurveillance systems is going to need to churn through massive amounts of diverse information and then present it to users in a way that can drive decisions and maintain some kind of situation awareness, either regional or national."

Some systems are already doing that by text mining online information streams to look for early warning signs of an outbreak. HealthMap, for example, mines close to 20,000 web pages to map the state of infectious diseases worldwide. BioCaster goes a little deeper, mapping layman's words for technical disease terminology in news reports in multiple languages (including several from Southeast Asia), looking to spot unusual trends, Kass-Hout says.

Kass-Hout helped develop another program, called Riff, while working at InSTEDD, a nonprofit international NGO founded by Google in 2006. This open source social networking platform scrapes information from RSS feeds or other Internet sources and uses a support vector machine behind the scenes to tag relevant items. As they are tagged, an expert gives a thumbs up or down to help the machine learn from its own mistakes.




***Polgreen's maximum coverage algorithm helped Iowa public health officials determine where to put new infectious disease surveillance sites. This map shows the 20 existing sites (pink) and 10 additional sites (green) identified by the maximum coverage calculator. The radius of coverage is 20 miles. The site shown in yellow covers a population of 158,571 people. Reprinted from P.M. Polgreen, et al., Optimizing Influenza Sentinel Surveillance at the State Level,* American Journal of Epidemiology, *170 (November 2009), pp 1300–1306, by permission of Oxford University Press.***

"Consensus develops and Riff does better over time," Kass-Hout says.

Social media sites also offer potentially valuable information, Polgreen says. His team used sixty million craigslist personals ads to study risk factors for sexually transmitted diseases. Because each listing is linked to a geographic location, his group can see how risky behaviors correlate with actual disease prevalence at a county level. Polgreen and his colleagues are also studying the Twitter stream to anticipate disease activity over the course of the flu season.

Remote sensing data can also help predict disease outbreaks in places where surveillance is difficult, such as in Southeast Asia. For example, researchers are using computational methods to interpret high-resolution images of environmental changes that could cause malarial outbreaks by boosting the mosquito population.

## Identifying Important Signals

Regardless of data input, biosurveillance systems must identify incidents that matter. As Kass-Hout puts it, users need to be able to ask: "What's happening here that needs my decision?" To that end, he has been working with collaborators to apply change-point analysis, traditionally used in econometrics, to the problem of determining when disease activity is stable, on the rise, or going down. "An early detection algorithm can't tell you that except when it goes above a certain background level." Hopefully, he says, the approach will complement traditional detection methods by identifying which signals need to get attention. The collaborators plan to publish their results soon.

## Making a Difference for Public Health

It's important to develop methods that are cutting edge but they must also be useful and understandable, Kass-Hout notes. "We don't want to build a highway to nowhere."

One system that's had some success in Illinois is called Indicator. It incorporates diverse data sources, can identify meaningful events, and has proven useful to public health officials, says **Ian Brooks, PhD**, director of the health sciences group at the National Center for Supercomputing Applications at the University of Illinois and one of the system's developers. Indicator can handle school district attendance information, hospital data, patient calls to advice nurses, and even veterinary surveillance. And it is flexible enough to allow the use of varied modeling approaches. "It's really a framework," Brooks says. So instead of committing to a particular model or way of determining whether the data is normal or aberrant, researchers can apply various algorithms to the data as it comes in, quickly creating models that are linked to that data. "You see an outbreak, and you can model how it will spread and how you would slow it down," Brooks says. During the H1N1 pandemic, he says, when school attendance data showed which schools were getting hit, the data gathered in Indicator suggested how the virus was spreading through the schools. "People then made decisions about vaccination strategies based on what we were seeing," Brooks says. "That's exciting."

> Polgreen's infectious disease work uses futures markets, data from social media sites, and iPhone apps. "I call it freakidemiology," he says.


***Screenshots from an iPhone app developed by Polgreen's team to help monitor hand hygiene in hospitals. Courtesy of Philip Polgreen.***

But to be useful for biosurveillance, a computational tool can be even simpler than that. Polgreen's team has built two tools that are easy to use by various different participants in the biosurveillance enterprise. One uses a maximal coverage algorithm to help public health officials decide where to locate their outpatient surveillance systems much as a retail company decides where to locate its next outlet. The second is an iPhone app for monitoring hand hygiene in hospitals. It replaces the pen, paper and clipboard surveillance currently in place in many hospitals to check whether healthcare workers are using hand hygiene as appropriate. Polgreen says it has already been downloaded by several thousand hospitals around the country. Projects like these, he says, provide valuable experience for graduate students while also making a difference in the real world.

## Forecasting with Freakidemiology

Polgreen also applies ideas from economics to forecast infectious disease spread. This approach was conceived by Polgreen after meeting the creators of the Iowa Electronic Markets (IEM), which have been forecasting election outcomes by using real money to trade in political futures since 1988. Polgreen and his colleagues then launched the Iowa Electronic Health Markets (IEHM), which have been used to forecast aspects of seasonal flu, avian flu and H1N1. For example, for H1N1, the market asked what the mortality rate will be or how many states the disease will spread to in the next month. People with prior experience of the disease are invited to participate in the market. They answer the question using their personal experience in the clinic and what they know about the disease. "It's like a survey on steroids," Polgreen says. "Ordinarily, it's very difficult to quantify subjective information but prediction markets help do that."

So far, he says, markets are producing pretty good predictions of flu spread. And although futures markets won't replace traditional forms of surveillance anytime soon, Polgreen says, when people have to make decisions with very little information, perhaps a little bit of freakidemiology can help. □