

Science Spotlight

June 2015

Using the Intersection of Statistics and Biology to Understand Flu Outbreaks

By Sara Mishamandani

A new method of data assimilation and surveillance has been developed by a research team at Columbia University to improve what we know about the dynamics of influenza transmission and to monitor the factors that can determine the intensity of a flu season. This novel approach, funded in part by the National Institute for Environmental Health Sciences (NIEHS), estimates epidemiological parameters that describe specific characteristics of influenza outbreaks. Using this methodology, public health officials can gauge the magnitude of an influenza threat and devise effective prevention and control measures.

Epidemiological parameters like individual susceptibility and rates of secondary infections, which are in turn related to frequent viral mutations, can alter the risks of major global pandemics. Influenza transmission is also strongly influenced by climate, with highest rates of spread in temperate climates occurring during cold, dry seasons. This study uses influenza as a model, improving understanding of how infectious diseases spread globally.

“By determining the key differences in influenza’s epidemiological characteristics from year to year, we can see how the characteristics of the influenza virus evolve over time, and how these changes lead to different epidemic intensities for different flu seasons,” said [Jeffrey Shaman](#), Ph.D., associate professor of environmental health sciences at Columbia University Mailman School of Public Health, and leader of the research team. “A better understanding of epidemiological parameters is also needed for generating better forecasts.”

A statistical method to understand and predict influenza patterns

Infectious disease surveillance systems are powerful tools to monitor infectious diseases, however, underreporting and observation errors can create challenges to fully understand influenza dynamics. In this study, the researchers used data from the Centers for Disease Control and Prevention (CDC) and from [Google Flu Trends](#), a Web service that aggregates Google search queries to make predictions about flu activity. Combining data from 115 cities over a ten-year period and using probability statistics, the team developed estimates of epidemiological characteristics of influenza outbreaks and variations of transmission among U.S. populations in different regions. The researchers also found that their approach compensated for observational errors and underreporting that characterize surveillance data.

The research team focused on three important epidemiological parameters related to influenza outbreaks. To determine how efficiently the virus can be transmitted from one person to another, they estimated the average number of secondary cases generated from a primary case in a susceptible population (referred to as the reproductive number). They also estimated the initial susceptibility of the population to determine the strength of immunity to the virus from year to year, and they developed a direct estimate of the burden of disease, known as the attack rate.

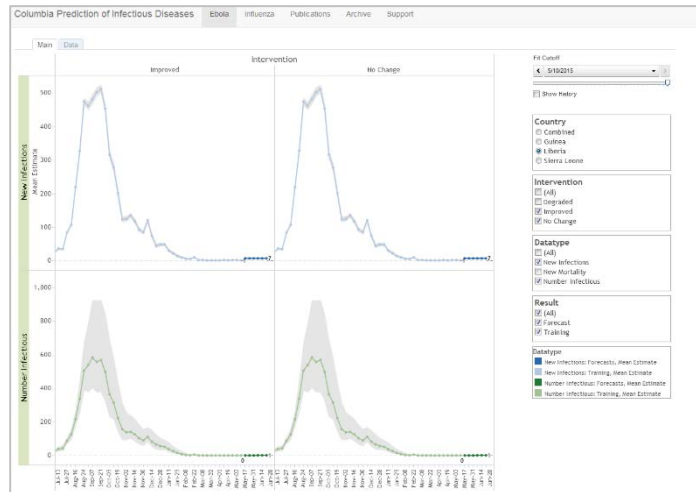
Their model demonstrated regional variations in influenza transmission dynamics. In particular, cities in the desert Southwest had basic reproductive numbers that did not correlate with the majority of other cities, suggesting that climate conditions may influence influenza transmission.

In addition to telling us more about the characteristics of influenza, the researchers are using the model to improve their infectious disease forecast system, which is available on the [Columbia University website](#).

“The statistical model-inference system applied in this study is used to optimize, or ‘train,’ our forecast system to generate a forecast,” said Wan Yang, Ph.D., an associate research scientist at Columbia and the lead author on the study. “It is the accurate estimation of the epidemiological parameters that provides the initial conditions for a forecast.”

Big data to improve disease surveillance

As technological capabilities allow us to collect and process larger and larger amounts of data, big data, a catch-phrase used to describe a volume of information that is so large it is difficult to process using traditional [database](#) and [software](#) techniques, is becoming part of every scientist’s vocabulary. Shaman and his research team are taking advantage of one of these huge datasets to improve their forecasts. The big data estimates of flu incidence from Google Flu Trends provided influenza activity data based on Google searches at the municipality level, which is more specific than regional data from the CDC. This allowed the researchers to study the spatial variation across the United States more precisely than with previous datasets.



During the cold and flu season Shaman’s group produces real-time forecasts of influenza for the U.S. at the municipal and state level. In response to the current Ebola outbreak in West Africa, they are now also producing real-time forecasts of this epidemic. (Photo courtesy of the Columbia Prediction of Infectious Diseases website)

According to the authors, as more people have access to and increasingly rely on online systems worldwide, mining of similar big data estimates from online social networks may provide valuable information on the early spread of infectious diseases as well as transmission dynamics for a number of other diseases, such as Ebola.

"These big data sources cannot replace traditional disease surveillance systems, but they could be an important supplement to the traditional systems and have been tremendously helpful for our work," said Shaman. "There are potential errors and biases in these big data sources so we look forward to further improvement of these systems and will continue to use them for our work."

"We are expanding our influenza inference systems to other regions around the world, for instance, subtropical regions where influenza could circulate year-round, in contrast to the wintertime flu epidemics typically observed in temperate regions," said Yang. "We are also working to expand our influenza model-inference systems and forecast systems to other infectious diseases such as Ebola and respiratory syncytial virus infection, another important respiratory infectious disease."

Citation:

Yang W, Lipsitch M, Shaman J. 2015. Inference of seasonal and pandemic influenza transmission dynamics. *Proc Natl Acad Sci* 112(9):2723-2728.[\[Abstract\]](#)