Flu season forecasts could be more accurate with access to health care companies' data

September 19 2018

New research shows that data routinely collected by health care companies could enable more accurate forecasts of when the next flu season will peak, how long it will last and how many people will get sick. Credit: University of Texas at Austin

In an era when for-profit companies collect a wealth of data about us, new research from The University of Texas at Austin shows that data collected by health care companies could—if made available to

researchers and public health agencies—enable more accurate forecasts of when the next flu season will peak, how long it will last and how many people will get sick.

In the U.S., seasonal influenza causes thousands of deaths and hundreds of thousands of hospitalizations each year. Forecasting can improve prevention, planning and care to reduce the human toll of severe seasonal and pandemic influenza.

Researchers for years have developed computer models for forecasting what an upcoming flu season will be like, but their results are often not very accurate. One major challenge is choosing the right kinds of data to feed into the models.

Professor Lauren Ancel Meyers and postdoctoral researcher Zeynep Ertem have developed a method for evaluating hundreds of data sets to find which are the most predictive and how to combine them to get the most accurate forecasts. In mathematical parlance, this is called an optimization problem.

Of the more than 600 flu-related data sets they evaluated, they found that some of the best predictions came from electronic health records collected by athenahealth, a company that provides cloud-based services for health care providers. These data, collected across the U.S., included information such as how many patients receive flu vaccinations, positive flu test results and flu-related prescriptions. Combining athenahealth's data with traditional surveillance data collected by the Centers for Disease Control and Prevention (CDC), which are still the best standalone data for predictions, would improve forecasts. The predictions were 15 percent more accurate with these combined data sets than if only CDC data were used.

Although athenahealth's data was provided to The University of Texas at

Austin for research purposes, it's difficult for researchers or [public health](#) agencies to access similar data from [health](#) care companies on an ongoing basis. The data is considered proprietary, and Meyers speculates that privacy issues would have to be worked out along with any expenses.

"Our study suggests it might be worth trying to cross some of those hurdles because the data can be quite powerful," Meyers said.

They published their results in the journal *PLOS Computational Biology*.

"Our method can be applied to any geographic region and to many other infectious diseases, including mosquito-transmitted viruses such as dengue and chikungunya," said Ertem.

The researchers found that the most predictive data sets were traditional surveillance sources collected from across the U.S. by the CDC. One, which the CDC calls ILINet, tracks weekly counts of patients seeking care for influenza-like illness, as reported by a sample of [health care](#) providers. The other collects data from more than 400 clinical labs across the U.S. and tracks the percentages of respiratory specimens that test positive for influenza.

Meyers said she hopes other researchers who are developing disease forecasting tools will apply these insights and their new methodology to improve the accuracy and timeliness of predictions.

"The message is that we should think more systematically about the data that fuel our disease forecasts," Meyers said. "With powerful—and sometimes surprising—combinations of data, we can make earlier and more accurate predictions about emerging threats."

More information: Zeynep Ertem et al, Optimal multi-source

Provided by University of Texas at Austin