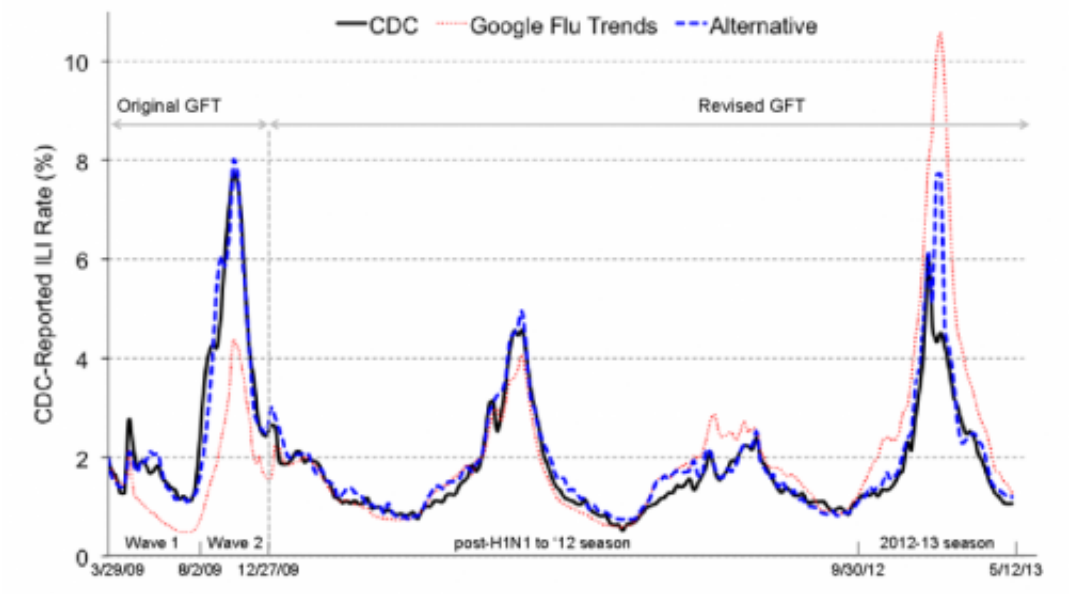
Finding real value in big data for public health

July 2 2014



For example, during the 2012/2013 season, GFT predicted that 10.6% of the population had influenza like illness when only 6.1% did according to patient records. The team's alternative significantly reduced the error in that prediction, estimating that 7.7% of people would have the flu. And within two weeks the model self-updated, considerably changing the weight given to certain queries that spiked during that time, improving the model for future performance. Credit: SDSU

Media reports of public health breakthroughs made possible by big data have been largely oversold, according to a new study, published today in the *American Journal of Preventive Medicine*.

"Many studies deserve praise for being the first of their kind, but if we actually began relying on the claims made by big data surveillance in public health, we would come to some peculiar conclusions," said John W. Ayers, San Diego State University Graduate School of Public Health research professor and senior author of the study. "Some of these conclusions may even pose serious public health harm."

But don't throw away that data just yet.

The authors maintain that the promise of big data can be fulfilled by tweaking existing methodological and reporting standards. In the study, the Ayers and his colleagues demonstrate this by revising the inner plumbing of the Google Flu Trends (GFT) digital disease surveillance system, which was heavily criticized last year (see here and here) after producing erroneous forecasts.

"Assuming you can't use big data to improve public health is simply wrong," added Ayers. "Existing shortcomings are a result of methodologies, not the data themselves."

A solution for Google Flu Trends

In the first external revision proposed to GFT, Ayers and co-researchers David Zhang, Maurcio Santiliana (both with Harvard University), and Benjamin Althouse (with the Santa Fe institute) explored new methods for using open-sourced, publicly available Google search archives to forecast influenza, an approach that can serve as a blueprint to fix broader shortcomings in public health surveillance.

To address GFT's problems, the team significantly beefed up the existing GFT model. First, rather than relying on a single trend that represents a group of influenza search queries, they monitored changes in individual search queries, giving various algorithmic weight to some queries over

others based on how they potentially improved predictions compared to patient data collected by health agencies.

Second, instead of relying on investigator opinion for periodic updates to the model, the team built in automatic updating that adjusts the weight given to any single query in the model each and every week based on artificial intelligence techniques to maximize predictive accuracy.

During the 2009 H1N1 pandemic and 2012/13 season—two critically important periods of influenza surveillance in the United States—the alternative method yielded more accurate influenza predictions than GFT every week, and was typically more accurate than GFT during other influenza seasons.

"With these tweaks, GFT could live up to the high expectations it originally aspired to," Ayers said. "Still, the greatest strength of our model is how the queries being used to describe influenza trends are changing over time as search patterns change in the population or the model occasional underperforms due to false-positive queries."

For example, during the 2012/2013 season, GFT predicted that 10.6% of the population had influenza like illness when only 6.1% did according to patient records. The team's alternative significantly reduced the error in that prediction, estimating that 7.7% of people would have the flu. And within two weeks the model self-updated, considerably changing the weight given to certain queries that spiked during that time, improving the model for future performance.

What's next for big data

"Big data is no substitute for good methods, and consumers need to better discern good from bad methods," Ayers said. To achieve these ends, he and his colleagues added that digital disease surveillance

researchers need greater transparency in the reporting of studies and better methods when using big data in public health.

"When dealing with big data methods, it is extremely important to make sure they are transparent and free," co-author Althouse added. "Reproducibility and validation are keystones of the scientific method, and they should be at the center of the big data revolution."

Importantly, these criticisms shouldn't be taken as an indictment of the promise of big data, or of the early attempts to wrangle it into something beneficial for the public, Ayers said. Now that the initial hype is wearing off, researchers can begin seriously exploring and testing the strengths and limitations of existing models and sharpening their methodologies.

"We certainly don't want any single entity or investigator, let alone Google—who has been at the forefront of developing and maintaining these systems—to feel like they are unfairly the targets of our criticism," Ayers said. "It's going to take the entire community recognizing and rectifying existing shortcomings. When we do, big data will certainly yield big impacts."

Provided by San Diego State University