How big data can enhance medical research

April 14 2016, by George Thomas Jr.

There's a reason "flu season" has earned its miserable prominence: When the flu is severe, it's difficult to avoid.

And if you're a parent you know this all too well, since children play a key role in influenza transmission.

That's why, in 2013, health officials in the United Kingdom selected children in seven cities across England for an experiment on the effectiveness of flu vaccines.

But when they tried to test the success of this campaign, they faced one critical hurdle: The flu season was not considered severe enough to provide definitive test results.

Or so they thought.

"What they were counting was how many people were hospitalized or visited the doctor," Microsoft principal researcher Elad Yom-Tov said. "But very few people who have the flu ever bother to see a doctor. Usually when you get the flu you stay at home a few days and that's that."

Instead, Yom-Tov and his colleagues at University College London looked at flu-related searches on Bing, along with how many people complained about the flu on Twitter.

"Based on these data sources alone, we were able to show a 25 to 30

percent reduction in the number of [flu cases](#) in cities where the vaccine was distributed, compared to other cities where it was not," he said.

In his new book, [Crowdsourced Health: How What You Do on the Internet Will Improve Medicine](#), Yom-Tov, a principal researcher based in Microsoft's New England lab, shows how the Internet's trove of data on how we research our own health can itself be mined for answers medical researchers otherwise have no way of getting.

To protect users' privacy, Yom-Tov used rigorous safeguards, such as examining data that had been anonymized and aggregated.

"Even when pulling data from Twitter, we're not identifying any of the people individually—we don't care who they are," Yom-Tov said of the flu study. "All we care about is how many people have the flu in those seven cities where the vaccine was used, compared to other cities."

He applied the same data mining approach to try to get a better understanding of the [side effects of prescription drugs](#).

Yom-Tov, in collaboration with Evgeniy Gabrilovich, analyzed search engine logs for associations between drugs and their possible side effects, and they discovered some side effects were being overlooked.

"We were missing side effects people don't realize are connected to the drugs because they're more benign or take a longer time to appear," he said. "This is due to the fact that traditional methods rely on people reporting such associations."

Yom-Tov said this deep data dive is the best way to detect such [side effects](#). And it can be done without compromising user privacy because the data is sourced from very large populations.

"We're not asking if John Smith asked about this drug or this side effect," he says. "What we're asking is how many people asked about this side effect after asking about this drug?"

Still, Yom-Tov acknowledges Internet data does not hold all the answers.

"It's worth saying that these data are not the magic bullet—they won't replace traditional ways of doing medical research."

Yom-Tov's hope is that the medical community will start collaborating more closely with computer scientists and think of Internet data as another source of information for their research—especially for data that's difficult or impossible to get in another way.

Although Internet search data can be a valuable source of data for medical researchers, Yom-Tov also cautioned that much of the medical information published online "is not of the highest quality."

He recommends that people rely on trustworthy sources like the Mayo Clinic, government organizations, and websites certified by the Health on the Net Foundation, which promotes useful and reliable online health information.

"Be careful about what kind of authority you attribute to the sources of information you find online," he said, noting it is a topic of upcoming research.

Provided by Microsoft